

ПОИСК ТЕЛЕГРАМ-КАНАЛОВ НА ЗАДАННУЮ ТЕМАТИКУ

Рыбаченко И.А.

Томский Политехнический Университет, ИШИТР, АЗ-36, e-mail: iar12@tpu.ru

Аннотация

В статье предлагается алгоритм для поиска телеграм-каналов, отражающих заданную тематику. Алгоритм принимает на вход начальный список релевантных телеграм-каналов, а также набор релевантных и набор нерелевантных статей Википедии. Результатом работы алгоритма является расширенный список релевантных телеграм-каналов. Демонстрируется работа алгоритма на примере поиска телеграм-каналов на тему Первой мировой войны. Исходный набор из 6 телеграм-каналов был расширен до 193 каналов. Для части найденных телеграм-каналов обнаружено несоответствие искомой тематике.

Ключевые слова: телеграм, релевантность, графовый анализ, снежный ком.

Введение

Телеграм является популярным средством массовой информации. Анализ телеграм-каналов может быть полезен, например, для оценки общественного мнения по определенным вопросам [1, 2]. Существует большое количество телеграм-каналов и в таком разнообразии может быть сложно найти сообщения, отражающие заданную тематику. Для поиска телеграм каналов может использоваться сервис TgStat, на котором собраны каналы на заданную тематику [3] или метод снежного кома [4]. В настоящей работе предлагается алгоритм для поиска релевантных телеграм-каналов. Алгоритм принимает на вход короткий список релевантных телеграм-каналов, а также набор релевантных и нерелевантных статей Википедии. Результатом работы алгоритма является расширенный список релевантных телеграм-каналов.

Описание алгоритма

Описываемый алгоритм состоит из двух последовательных этапов. На первом этапе формируется список ключевых слов специфичных для заданной тематики. Поиск ключевых слов сводится к задаче извлечения именованных сущностей. Для извлечения именованных сущностей используется библиотека stanza. Ключевые слова извлекаются из заданного набора релевантных и нерелевантных статей Википедии. Между найденными списками ключевых слов вычисляется разница.

На втором этапе выполняется поиск телеграм-каналов. Для поиска телеграм-каналов используется метод снежного кома. Список телеграм-каналов итеративно расширяется через просмотр переотправленных сообщений и переход к каналу-отправителю сообщения. Найденные телеграм-каналы фильтруются по признаку вхождения ключевых слов.

Далее в статье приводится более подробное описание реализации алгоритма и результаты его работы. В качестве примера рассматривается поиск телеграм-каналов на тему первой мировой войны.

Поиск ключевых слов

В качестве заведомо релевантной статьи из Википедии была взята статья под заголовком «Первая мировая война». В качестве заведомо нерелевантной статьи использовалась статья «Вторая мировая война». Набор релевантных и нерелевантных статей был расширен путем добавления ссылок из исходных статей. Просматривались только ссылки из введения статей. Исключались ссылки на общеупотребимые термины. Общее количество включенных релевантных статей составило 27, нерелевантных – 41.

Именованные сущности извлекались при помощи библиотеки stanza. Использовались следующие модели – tokenize=syntagrus, pos=syntagrus_charim, lemma=syntagrus_nocharim, depparse=syntagrus_charim, ner=wikiner. Извлечение именованных сущностей из релевантных статей заняло 41 минуту и было найдено 24586 сущностей. Обработка нерелевантных статей заняла 78 минут и было найдено 20495 сущностей. Компьютерные вычисления производились на облачной виртуальной машине со следующими характеристиками: платформа Intel Ice Lake, 8 vCPU, 32GB RAM.

На рис. 1 представлено распределение найденных именованных сущностей по категориям. Большинство найденных сущностей являются географическими локациями. Названия сражений и мирных договоров попали в категорию прочее.



Рис. 1. Тематика найденных ключевых слов

Найденные именованные сущности были приведены к нижнему регистру. Из списка именованных сущностей про Первую мировую войну были исключены сущности содержащиеся в списке про Вторую мировую войну. В результате осталось 13842 сущности.

Одна именованная сущность могла встречаться в тексте несколько раз. После удаления дублирующих записей осталось 7346 сущностей.

В списке сущностей были обнаружены короткие малозначащие последовательности символов, например «дж», «цсу», «торн». После удаления сущностей с длиной менее 5 символов осталось 7094 сущности.

Среди найденных именованных сущностей имеются повторения и чрезмерные обобщения, например «кусманек» - «кусманека» (Герман Кусманек фон Бургнойштедтен, австро-венгерский генерал-полковник), «1-й гвардейский» - «гвардейский» (подразделение российской армии), «августовская операция» - «операция» (наступление 1915 года в районе города Августов). Также среди найденных именованных сущностей имеются такие общие слова как «репарации», «русская», «солдаты», «граната».

Поиск телеграм-каналов

Для расширения списка телеграм-каналов использовался метод снежного кома, изображенный на рис. 2. Для каждого необработанного телеграм-каналы выгружались первые 1000 сообщений. Для перепроверенных сообщений канал-источник анализировался на релевантность. Релевантность оценивалась по первым 1000 сообщений из телеграм-канала. Для каждого сообщения вычислялось сколько в нем содержится различных ключевых слов из ранее составленного списка. Если нашлось 5 и более сообщений, содержащих 5 и более ключевых слов, то телеграм-канал считался релевантным. Новые найденные релевантные каналы добавлялись в список необработанных. Процесс повторялся итеративно.

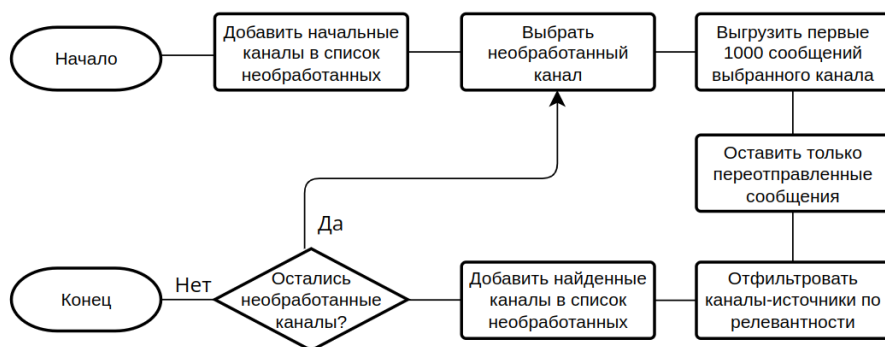


Рис. 2. Алгоритм поиска телеграм каналов

Для взаимодействия с телеграм использовалась библиотека telethon. Между вызовами api добавлялась задержка в одну секунду. Для оптимизации количества запросов через api использовался локальный кеш.

Изначальный список телеграм-каналов состоял из 6 элементов — @vanillakaiser, @wwimemes, @world_war_1, @diletant_media, @rf_history, @historyrussi. Обработка начального списка каналов заняла 12 минут и расширила количество каналов до 29. Обработка следующих 23 каналов заняла 87 минут и расширила количество каналов до 193. Третья итерация не была выполнена из-за проблем с сетью в процессе обработки.

На рис. 3 представлена визуализация найденного кластера каналов. Слабая связанность граничных вершин говорит о незавершенности итерационного процесса построения графа.

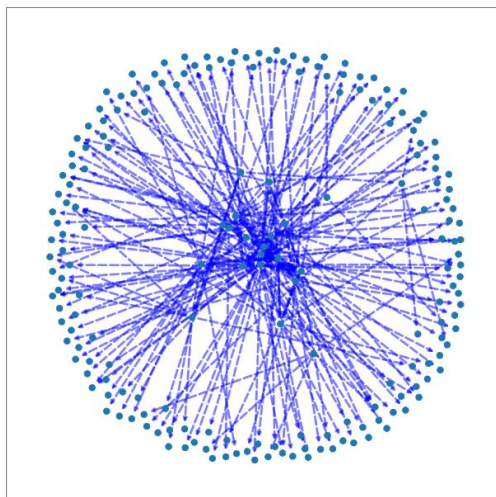


Рис. 3. Граф найденных телеграм-каналов

При анализе найденных телеграм-каналов было обнаружено, что часть из них посвящена Специальной военной операции на Украине (СВО). Чтобы этого избежать, можно было попробовать при поиске ключевых слов в качестве нерелевантных статей добавить статью про СВО.

Результаты

Предложенный алгоритм продемонстрировал эффективность в расширении списка телеграм-каналов, но показал проблемы с оценкой релевантности найденных телеграм-каналов. Из 66 просмотренных вручную телеграм-каналов, релевантными оказались 29, что составляет 43 %.

Список релевантных и нерелевантных статей Википедии был составлен вручную через просмотр ссылок из введения статьи на Википедии. Этот процесс можно было автоматизировать и добавить рекурсивный переход по ссылкам. Также при переходе по ссылкам можно было добавлять вес ключевых слов в зависимости от глубины перехода.

Используемый метод оценки релевантности телеграм-канала основывается на ручных правилах. Для повышения качества оценки релевантности можно было применить методы машинного обучения и подготовить более точный классификатор [5].

Заключение

В настоящей работе был продемонстрирован алгоритм для поиска телеграм-каналов на заданную тематику. На основе нескольких статей из Википедии и начального списка телеграм-каналов было найдено 193 телеграм-канала. Часть найденных телеграм-каналов посвящена Первой мировой войне, а часть – Специальной военной операции на Украине.

Список использованных источников

1. Габараева М.Р. Брачные отношения населения Республики Дагестан: контент-анализ онлайн-медиа / М.Р. Габараева // Kavkaz-Forum. – 2020. – № 3(10). – С. 148-160. – DOI 10.46698/P9405-6546-9208-f. – EDN TBQSCWW.
2. Милованов И.В. Возможности исследования настроений населения на основе новостных каналов социальной сети "Телеграм" / И.В. Милованов // Социальные практики и управление: проблемное поле социологии: Материалы VI сибирского социологического форума с международным участием, Новосибирск, 27 октября 2023 года. – Новосибирск: Новосибирский государственный университет экономики и управления "НИНХ", 2023. – С. 275-279. – EDN HEMNLF.

3. Углова А.Б. Анализ деструктивного контента телеграмм-каналов как фактора развития саморазрушающего поведения / А.Б. Углова, Б.А. Низомутдинов // International Journal of Open Information Technologies. – 2022. – Т. 10. – № 11. – С. 81-86. – EDN TAKUEF.
4. The Pushshift Telegram Dataset [Electronic resource] / J. Baumgartner [et al.] // Proceedings of the International AAAI Conference on Web and Social Media. – 2020. – Vol. 14. – P. 840–847. – Available from: <https://doi.org/10.1609/icwsm.v14i1.7348>.
5. Минаев В.А. "Просеивание" телеграм-каналов при поиске контента экстремистского характера / В. А. Минаев, А. В. Симонов // Информация и безопасность. – 2023. – Т. 26, – № 1. – С. 25-30. – DOI 10.36622/VSTU.2023.26.1.003. – EDN GONJUB.