ФГБОУ ВО «ЧелГУ»
Математический факультет
Кафедра компьютерной безопасности и прикладной алгебры

Дипломная работа

Сбор и анализ больших данных из социальных сетей

Автор:
студент группы МК-601
Кривощеков А.М.
Научный руководитель:
старший преподаватель
Фельдман В.В.

Большие данные

Характерные черты больших данных:

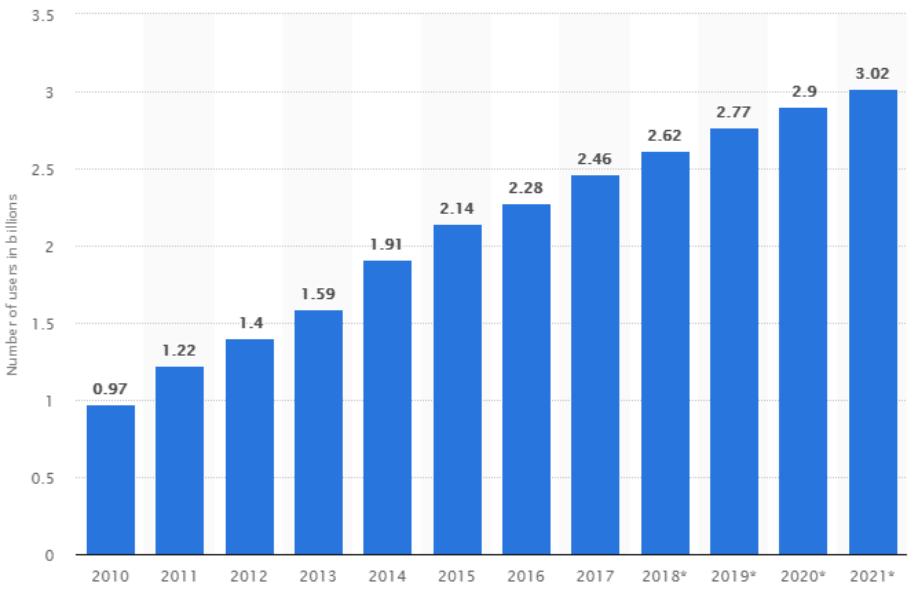
- Большой объем
- Разнообразие
- Изменчивая интенсивность

sas.com

В рамках данной работы, производится сбор и анализ больших объемов слабоструктурированных комментариев пользователей из социальных сетей.

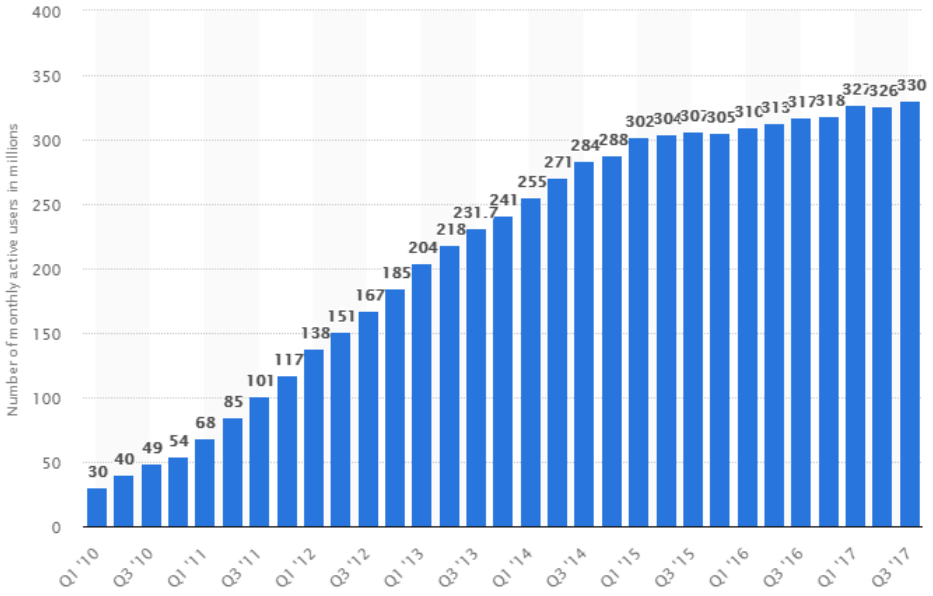
Актуальность

Количество пользователей социальных сетей по всему миру с 2010 по 2021 (миллиардов/год)



Источник: https://www.statista.com

Количество активных пользователей социальной сети Twitter (миллионов/квартал'год)



Источник: https://www.statista.com

Области и способы применения данных из социальных сетей

Область	Способ
Маркетинг	• Сбор отзывов о выпущенном продукте/услуге • Точный подбор заинтересованных потребителей продукта/услуги • Обзор и аналитика конкурентов и т.д...

Политика	• Выявление случаев нарушения законодательства:
	  • Выявление случаев экстремизма
	  • Выявление случаев неуважения чувств верующих
	  • Выявление случаев разжигания межнациональной розни
	и т.д...

Проблема исследования

заключается в повышении эффективности сбора и анализа больших данных в условиях отсутствия прямого неограниченного доступа к ресурсам социальных сетей

Цель исследования

разработать программный комплекс, обеспечивающий получение, хранение и определение тональности больших объемов слабоструктурированных комментариев пользователей из социальных сетей.

Задачи исследования

- изучить методы построения высоконагруженных систем хранения и обработки данных;
- разработать и апробировать программный комплекс, включающий веб-сервер контроллер, управляемый клиент (или их множество), базу данных (или их множество);
- проанализировать и описать результаты исследования, полученные при помощи алгоритма машинного обучения.

Подзадачи

Подзадача	Возможное решение
Большие объемы информации	Использование специализированных распределенных баз данных
Доступ к api осуществляется через интернет → временные задержки выполнения	Микросервисная архитектура

Для enterprise-проектов есть возможность приобретения лицензий на доступ к данным. Условия обговариваются индивидуально (прим. Twitter). Получение лицензии значительно увеличивает лимит ресурсов, или же снимает их совсем.

Подзадачи

Подзадача	Возможное решение
Большие объемы информации	Использование специализированных распределенных баз данных
Доступ к api осуществляется через интернет → временные задержки выполнения	Микросервисная архитектура

Для enterprise-проектов есть возможность приобретения лицензий на доступ к данным. Условия обговариваются индивидуально (прим. Twitter). Получение лицензии значительно увеличивает лимит ресурсов, или же снимает их совсем.

Хранение данных

Хранимые данные характеризуются следующим образом:

- Большой объем
- Множество полей
- Интересующие поля – преимущественно текстового типа (textgeneral)

Выбор был сделан в пользу программного продукта Apache Solr

- Высоконадежная
- Масштабируемая
- Отказоустойчивая
- Распределенное индексирование
- Репликация
- Балансировка нагрузки
- Автоматическое восстановление после сбоев
- Централизованная конфигурация

Подзадачи

Подзадача	Возможное решение
Большие объемы информации	Использование специализированных распределенных баз данных
Доступ к api осуществляется через интернет → временные задержки выполнения	Микросервисная архитектура

Для enterprise-проектов есть возможность приобретения лицензий на доступ к данным. Условия обговариваются индивидуально (прим. Twitter). Получение лицензии значительно увеличивает лимит ресурсов, или же снимает их совсем.

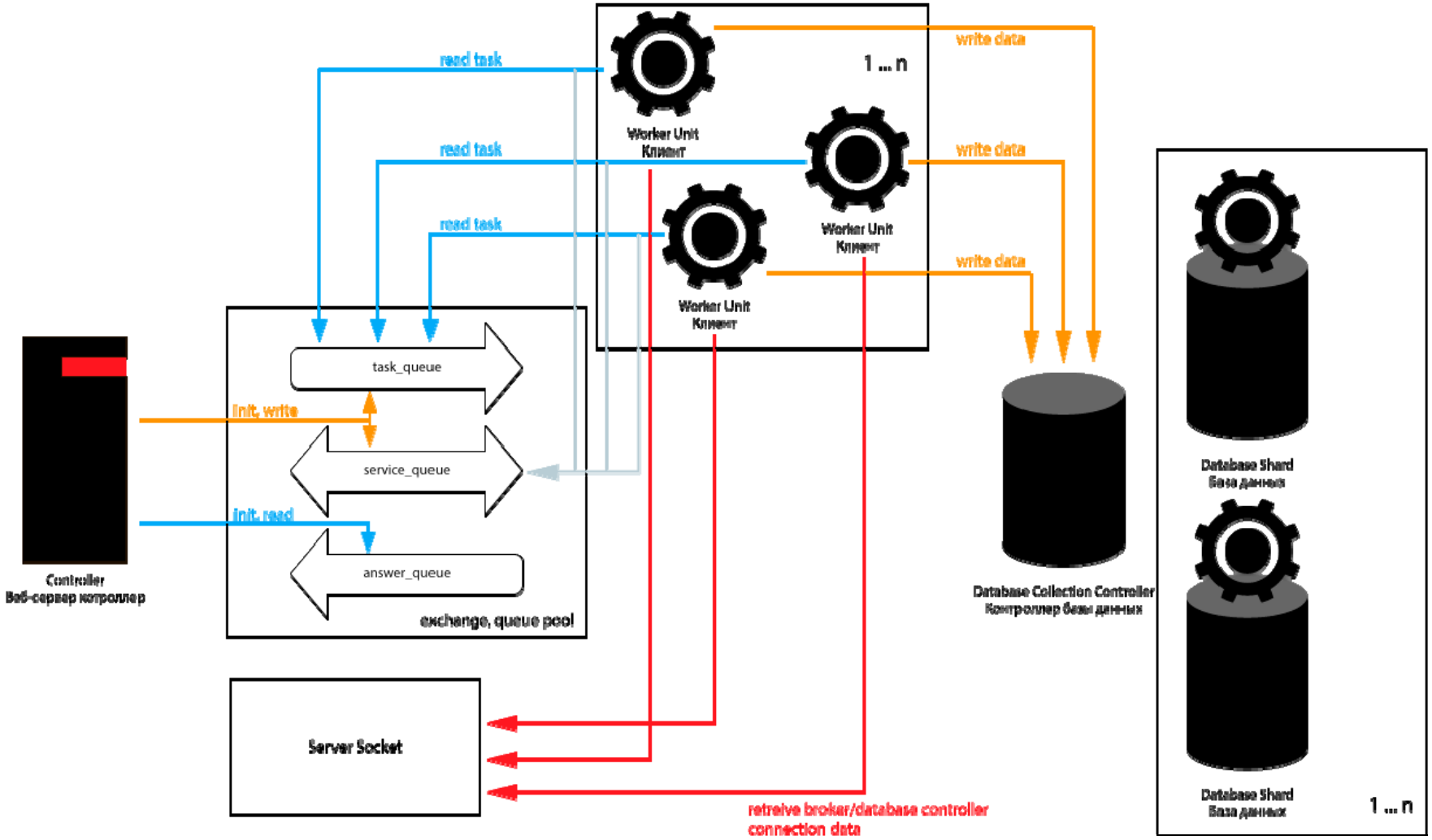
Микросервисная архитектура

Выбор в сторону архитектуры микросервисов (SOA)

Реализована схема, состоящая из следующих компонентов:

- Веб-сервер контроллер (Java)
  - Брокер очередей RabbitMQ – [ссылка]
  - Библиотеки для работы с брокером для Java
  - Локальная база MySQL для хранения конфигурации и служебных данных
  - Библиотеки для работы с MySQL для Java
- Клиент (PHP)
  - Библиотека Solarium для работы с базой данных Solr – [ссылка]
  - Библиотека для работы с Twitter-Api – [ссылка]
  - Библиотека для работы с RabbitMQ – [ссылка]
- Контроллер базы данных (Solr – [ссылка])
- Протокол обмена сообщениями через RabbitMQ между клиентами и веб-сервером контроллером

Схема системы получения и хранения данных



Обработка полученных данных

Было решено сделать выбор в пользу Apache Spark

Преимущества Apache Spark:

- Включает в себя библиотеку машинного обучения MLlib с множеством готовых к использованию методов
- Возможно получение данных напрямую из Solr через адаптер от разработчика

Анализ тональности

Метод основан на градиентном «бустинге» над решающими деревьями.

При обучении классификатора использовались вручную отобранные сообщения в количестве:

- Негативные – 110
- Позитивные – 73

Итоговая точность распознавания (суммарно, на обучающей и проверочной выборках):

- Негативных сообщений – 95%
- Позитивных сообщений – 82%

Классификатор будет улучшаться в дальнейшем

Заключение

В результате исследования был разработан программный комплекс, обеспечивающий получение, хранение и определение тональности больших объемов слабоструктурированных комментариев пользователей из социальной сети twitter.

В дальнейшем, планируется доработать и перенести комплекс на облачные ресурсы, добавить поддержку других социальных сетей, реализовать новые алгоритмы машинного анализа, упаковать проект в готовый сервис с привязкой веб-сайта.

Спасибо за внимание!