

Распределенная классификация текстового потока: проблемы, ограничения и решения

Шавкунов Михаил Валерьевич

научный руководитель: И. Е. Кураленок

НИУ ВШЭ

7 июня 2019 г.

Задача классификации новостного потока:

- На вход системе подается поток новостей
- Каждую новости нужно сопоставить подходящую тему
- Потенциально данных может быть бесконечно

Задача обобщается и на другие: выявление подозрительных транзакций, анализ сообщений пользователей и тп.

К таким задачам обычно выдвигаются следующие требования:

- Минимальная задержка: обработать каждый входной элемент как можно быстрее
- Масштабируемость: данных много и их невозможно обработать на одном компьютере

1. Использование библиотеки scikit learn:
 - Приемлемая точность¹
 - Нет масштабируемости
2. Системы пакетной обработки данных (MapReduce)
 - Масштабируемость и множество гарантий на данные
 - Высокая задержка вычислений для каждого элемента – до нескольких часов
3. Поточковые системы

¹F. Pedregosa и др. “Scikit-learn: Machine Learning in Python”.
В: *Journal of Machine Learning Research* 12 (2011), с. 2825—2830.

1. Корректность результатов: отсутствие отказоустойчивости
2. Низкая задержка вычислений при отсутствии гарантий на данные
3. Воспроизводимость: отсутствие детерминизма

Цель: реализовать фреймворк для потоковой классификации текстов и проверить влияние детерминизма на систему

Задачи:

1. Выбор платформы
2. Вычисление признаков для классификации
3. Дообучение классификации
4. Доработка модели
5. Проведение тестирования

Задача 1: выбор платформы

| Система | Детерминизм | Задержка |
|---------------------|-------------|------------|
| Storm, Heron, Samza | – | низкая |
| Flink | – | низкая |
| MillWheel | – | недоступно |
| Spark Streaming | + | высокая |
| FlumeStream | + | низкая |

Сравнение потоковых систем²

низкая задержка — до 100 миллисекунд

высокая задержка — больше 1 секунды

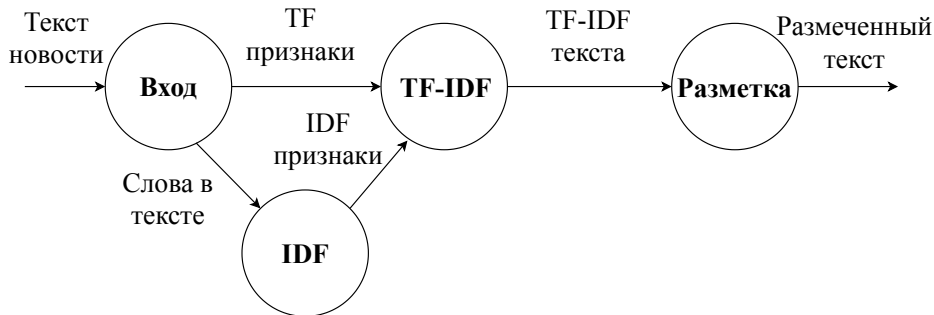
²Igor E. Kuralenok и др. “FlumeStream: Model and Runtime for Distributed Stream Processing”. В: *Proceedings of the 5th ACM SIGMOD Workshop on Algorithms and Systems for MapReduce and Beyond*. BeyondMR’18. Houston, TX, USA: ACM, 2018, 8:1—8:2. ISBN: 978-1-4503-5703-6. DOI: 10.1145/3206333.3209273. URL: <http://doi.acm.org/10.1145/3206333.3209273>.

Открытые данные с новостного сайта lenta.ru³:

- Реальные статьи за последние 20 лет
- 90 различных тем новостей

³*Lenta.ru dataset*. Февр. 2019. URL:
<https://github.com/yutkin/Lenta.Ru-News-Dataset>.

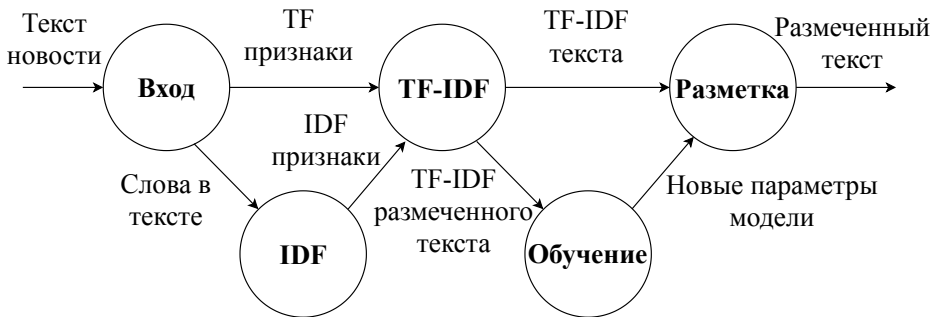
Задача 2: вычисление признаков



IDF как отдельная вершина:

- Зависимость вычисления признаков от всей коллекции
- Добавление текста в коллекцию документов

Задача 3: схема вычислений с обучением



Использование онлайн обучения:
размеченные тексты обновляют модель поточно

Задача 4: доработка модели

- Наличие стартовых параметров классификатора
- Небольшой размер модели для хранения и обновления на всех узлах
- Разумная скорость для обработки в реальном времени

Решение: использование L1 регуляризации для получения разреженных весов модели

Задача 5: тестирование

Проверка влияния недетерминизма:

| Размер кластера | % документов с различными темами ¹ (среднее \pm ср.кв.отклонение) | Точность ² % (среднее \pm ср.кв.отклонение) |
|-----------------|---|---|
| 2 | 0.9 ± 0.2 | 77.3 ± 0.2 |
| 4 | 1.7 ± 0.4 | 77.3 ± 0.2 |
| 8 | 1.9 ± 0.5 | 77.3 ± 0.2 |

¹ на FlameStream: 0 ± 0

² на FlameStream: 77.3 ± 0

10 независимых запусков 10 000 документов на Flink

Вывод: в следствие недетерминизма системы, с ростом размера кластера процент различных тем увеличивается

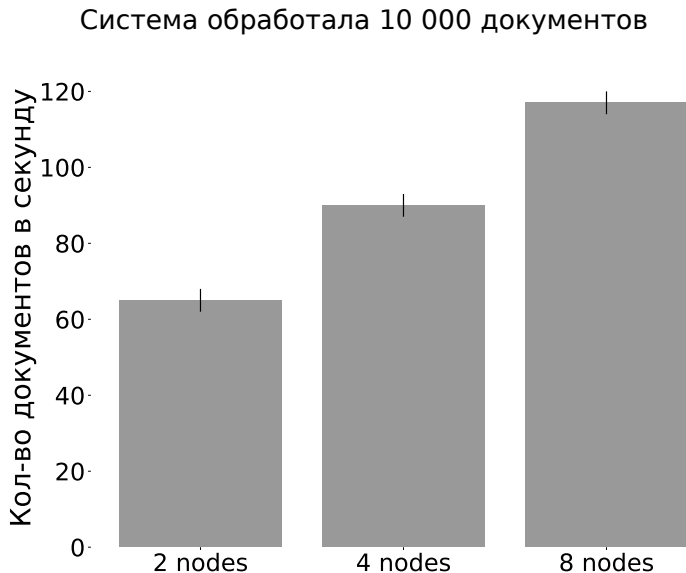
Сравнение с фреймворком Sklearn по качеству классификации:

| Метод | % Точность |
|-------------------------------|------------|
| Статическое обучение(sklearn) | 69 |
| Онлайн обучение на потоке | 68 |

В тренировочной выборке – 10 000 документов

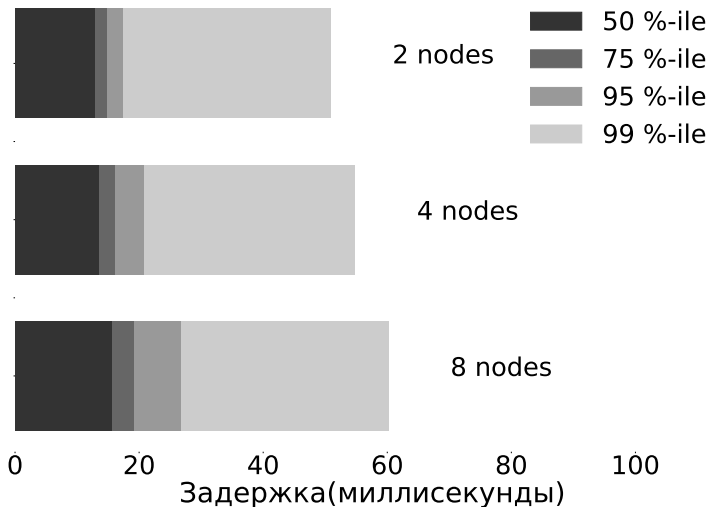
В тестовой – 5 000

Задача 5: тестирование



Задача 5: тестирование

Система обработала 10 000 документов



- Реализована классификация с онлайн дообучением – точность не ниже 1% по сравнению с Sklearn
- Фреймворк масштабируем и работает с низкой задержкой – до 100 мс
- Влияние недетерминизма – обнаружено отклонение до 2% в разметке новостей
- Публикация на конференции ACM DEBS