Применение машинного обучения для анализа использования API

Чебыкин Александр Евгеньевич, 371 группа

СПбГУ, Математико-механический факультет

Научный руководитель: ст. пр. кафедры системного программирования Я. А. Кириленко

Мотивация

- Программирование связано с использованием разнообразных библиотек
- Источник информации об их применении - чаще всего интернет
- Альтернатива: получение правил использования API из исходного кода
  - Частотный анализ
  - Машинное обучение - более мощные статистические модели

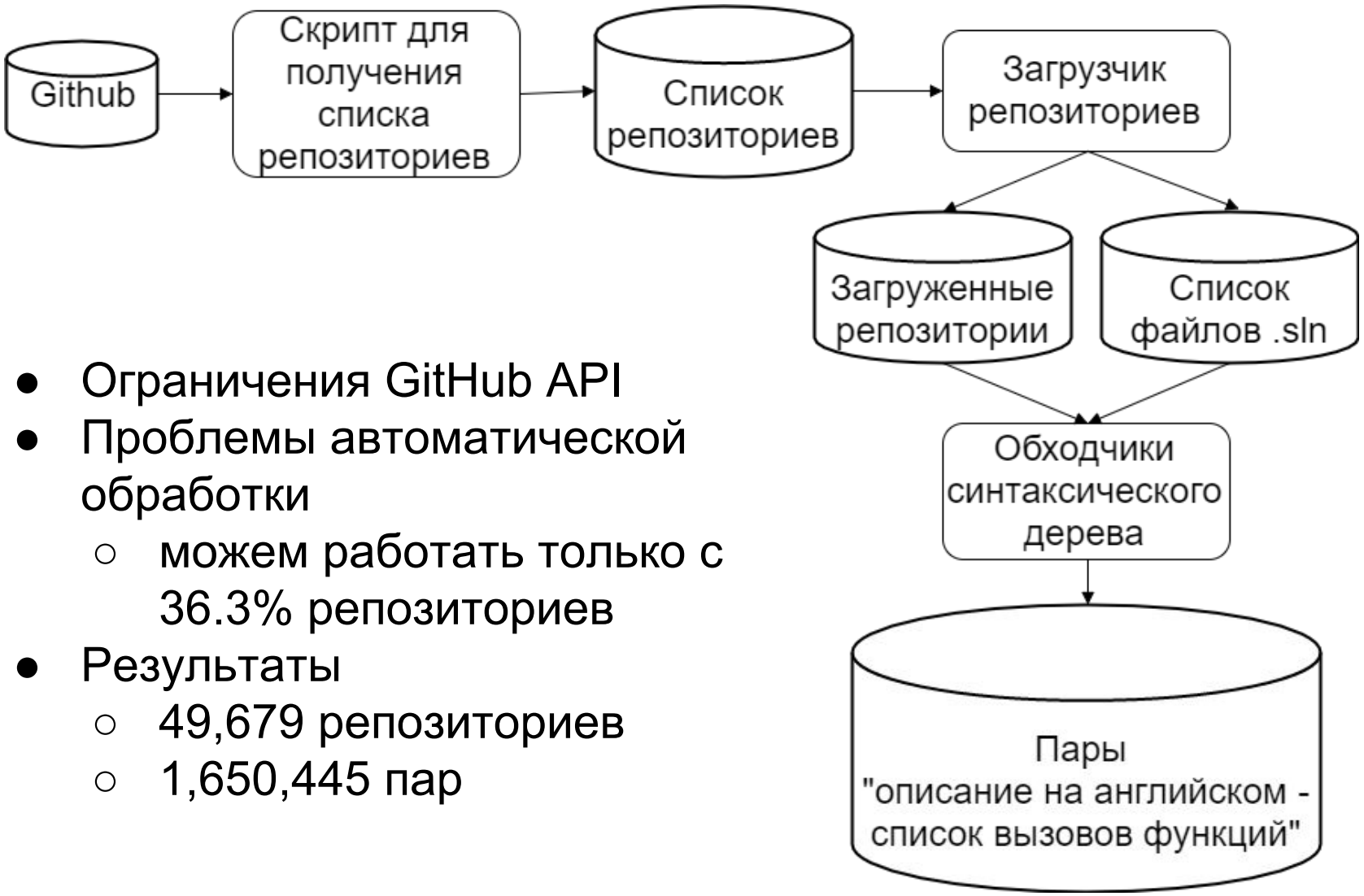
Deep API Learning

- Предложение цепочки API по текстовому запросу
  - “generate random number” – “Random.new Random.Next”
- Глубокое машинное обучение из области машинного перевода
- Sequence-to-Sequence
  - RNN Кодер-Декодер
- Существует публичный рабочий прототип
  - Релевантные ответы на запросы

Постановка задачи

- Реализация алгоритма на основе другого языка программирования
- Исследование проблем переноса на другой язык
- Исследование релевантности подхода

Получение данных для обучения



- Ограничения GitHub API
- Проблемы автоматической обработки
  - можем работать только с 36.3% репозиториев
- Результаты
  - 49,679 репозиториев
  - 1,650,445 пар

Извлечение данных

```
/// <summary>
/// Writes a <see cref="TimeSpan"/> value.
/// </summary>
/// <param name="value">The <see cref="TimeSpan"/> value to write.</param>
15 references | James Newton-King, 1157 days ago | 2 authors, 3 changes
public override void WriteValue(TimeSpan value)
{
    base.WriteValue(value);
    AddToken(new BsonString(value.ToString(), true));
}
```

Комментарий: writes a timespan value

Список вызовов функций: JsonWriter.WriteValue; TimeSpan.ToString; BsonString.new; BsonWriter.AddToken

Реализация модели

- TensorFlow
  - Ресурсы железа не позволяют запустить модель с оригинальными параметрами
  - Работает плохо, единичные результаты

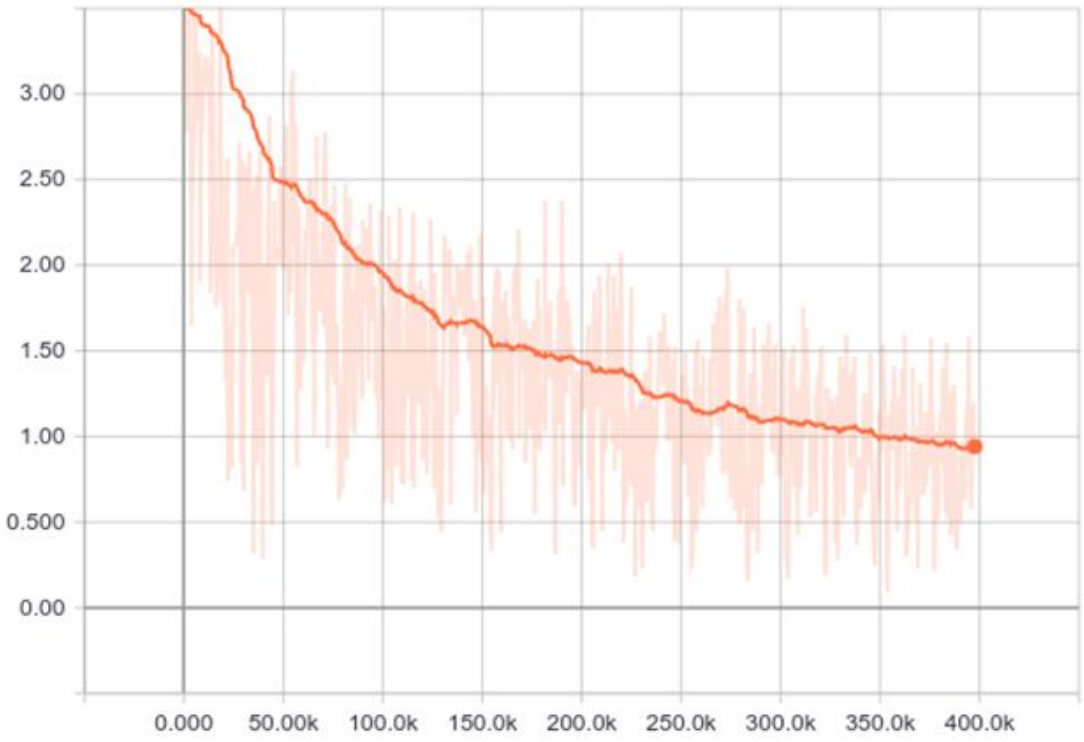
- tf-seq2seq
  - Свежая библиотека
  - Обучалась с оригинальными параметрами
  - 900 000 пар для обучения
  - Результаты есть, но хуже оригинальной статьи
    - Видимо, мало данных

Метрики

BLEU

Итерация	BLEU
50 000	0.00
158 000	0.48
258 000	1.03
390 000	1.95

Функция потерь



Итоги

- Проведен эксперимент по построению и обучению модели
- Реализованы инструменты для сбора данных
- Исследованы проблемы смены целевого языка