Детектирование смены автора в тексте

А.Фаттахов

Московский Физико-Тенический Институт

Москва, 2018

Цели исследования

- **Изучение методов поиска заимствований**
- **Улучшения качества классификации при «внутреннем» подходе**
- **Сравнение с другими подходами**

Литература

- Kuznetsov, M.; Motrenko, A.; Kuznetsova, R.; Strijov, V. Methods for IntrinsicPlagiarism Detection and Author DiarizationNotebook for PAN at CLEF 2016.
- han, J. Style Breach Detection: An Unsupervised Detection ModelNotebook for PANat CLEF 2017
- Safin, K.; Kuznetsova, R. Style Breach Detection with Neural Sentence EmbeddingsNotebook for PAN at CLEF 2017

Подходы к поиску заимствований

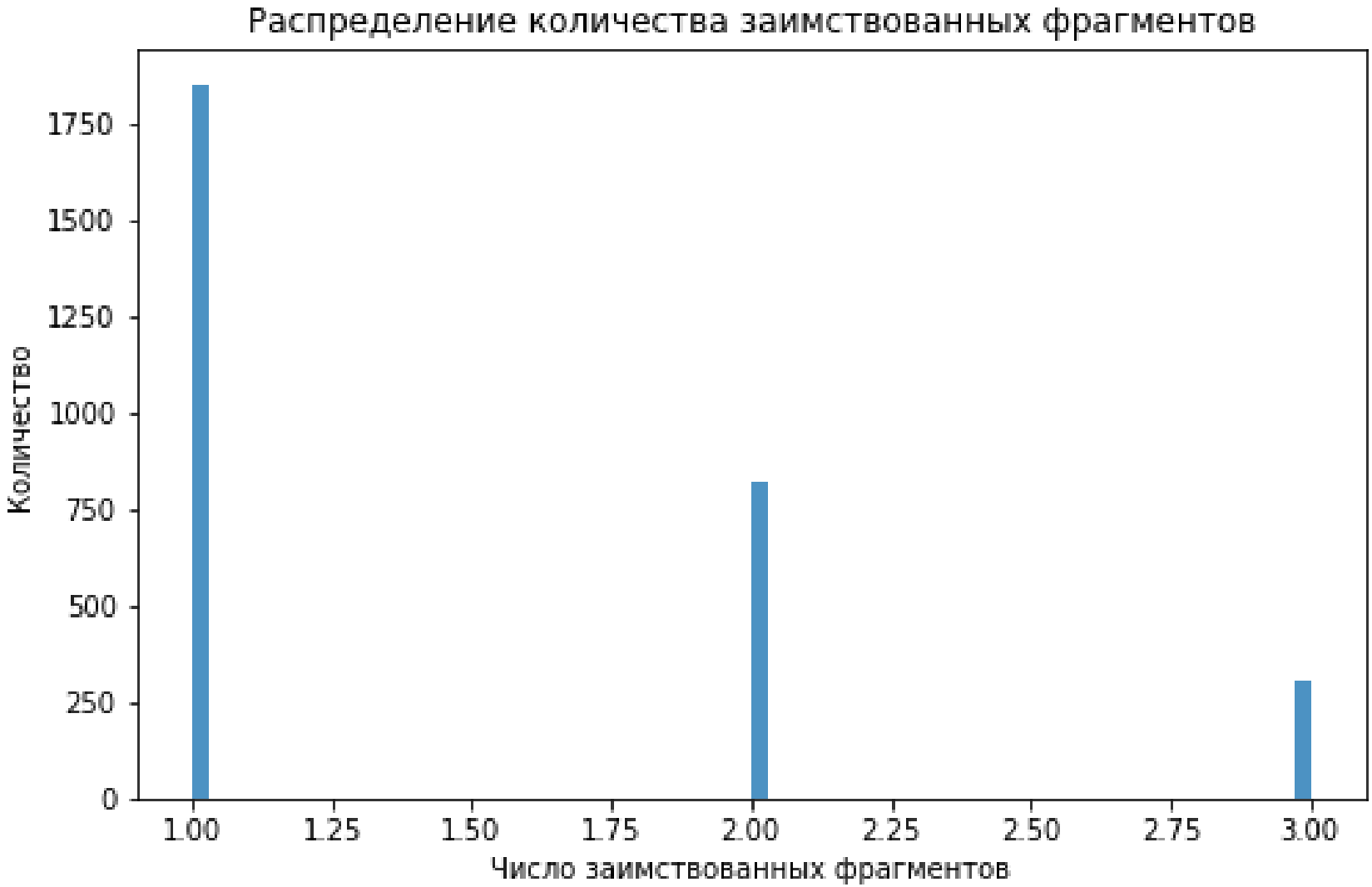
1. **«Внешний»**
   - Известен корпус, из которого производится заимствование
   - Можно опираться на внешние документы
2. **«Внутренний»**
   - Неизвестно корпуса
   - Нужно опираться исключительно на анализ исследуемого текста

Постановка задачи

Поставим формально задачу классификации текста. Дана выборка $D = \{(d_k; y_k)\}$, где d_k — текстовый документ, а $y_k \in Y = \{0, 1\}$ - является ли текст мультиавторным. Выборка разбита на обучающую D_l и контрольную D_k. Принята функция ошибки S. Требуется построить отображение $g : d_i \mapsto \mathbf{d_i}$ и модель классификации $f : R^n \mapsto Y$, минимизирующую функцию ошибки S на контрольной выборке:

$$f = \frac{1}{|D_k|} \sum_{1}^{|D_k|} \arg\min_f S(f(\mathbf{d_i}), y_i | D_l) \qquad (1)$$

Описание выборки



Описание выборки