Классификация авторства коллаборативных текстов

Выполнил: Рауф Курбанов
Руководитель: Алексей Шпильман

Актуальность

- Литература
- История
- Антиплагиат
- Научные публикации
- Политология

Постановка задачи

- Существуют произведения, написанные несколькими авторами: Братья Стругацкие, Ильф и Петров, Перумов и Лульяненко

- Задача: получать спектральную картину авторства таких произведений

Существующие решения

- Koppel, Moshe, et al. "Unsupervised decomposition of a document into authorial components."

- Akiva, Navot, and Moshe Koppel. "Identifying Distinct Components of a Multi-author Document."

- Tschuggnall, Michael, and Günther Specht. "Automatic Decomposition of Multi-Author Documents Using Grammar Analysis."

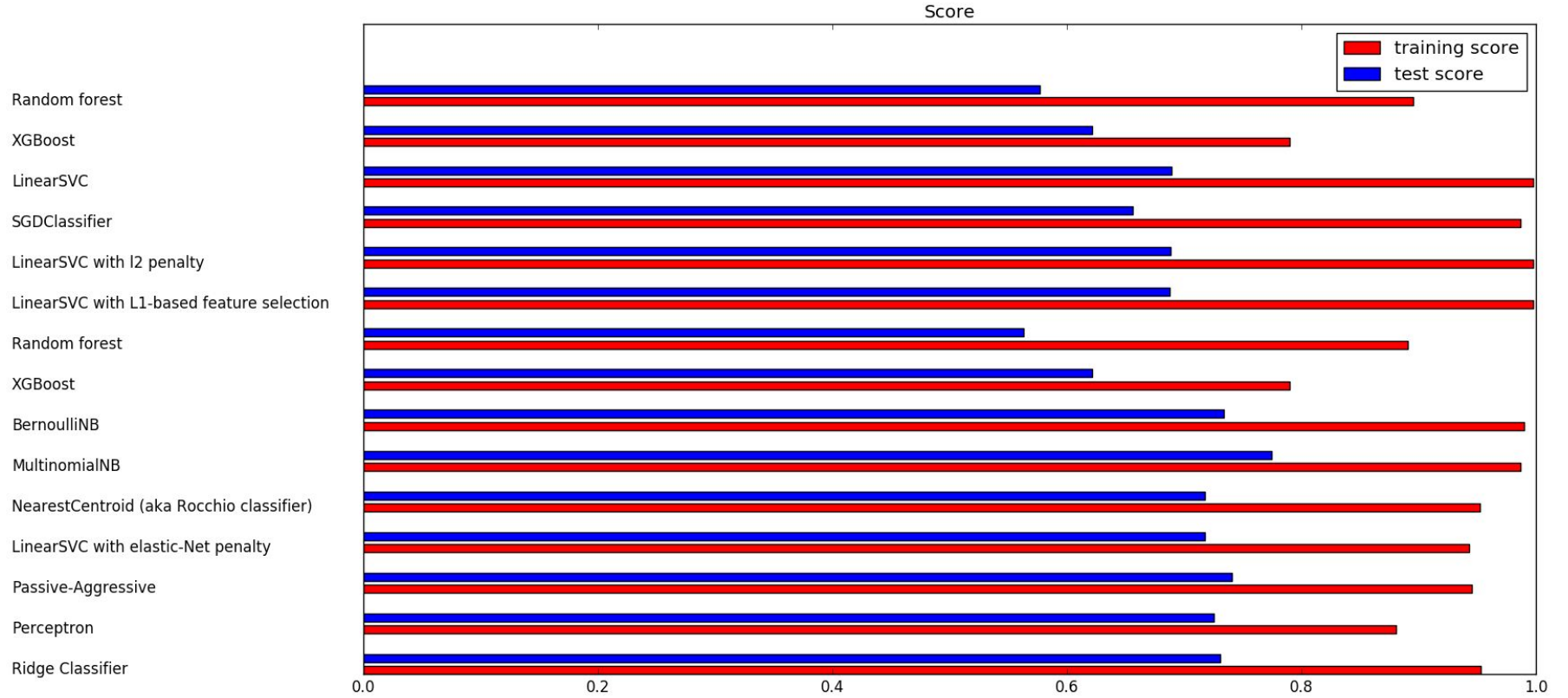
Фреймворк

- Feature engineering
  - Масштабируемая архитектура для создания набора фичей, как из готовых, так и предоставленные пользователем
  - Векторизация посредством bag of words
  - Более 20 специально разаработанных фич
- Глубокие нейросети с фиксированной геометрией
- Визуализация
  - Спектральная картина, подбор гиперпараметров, декомопзиция
- Кэширование и сериализация

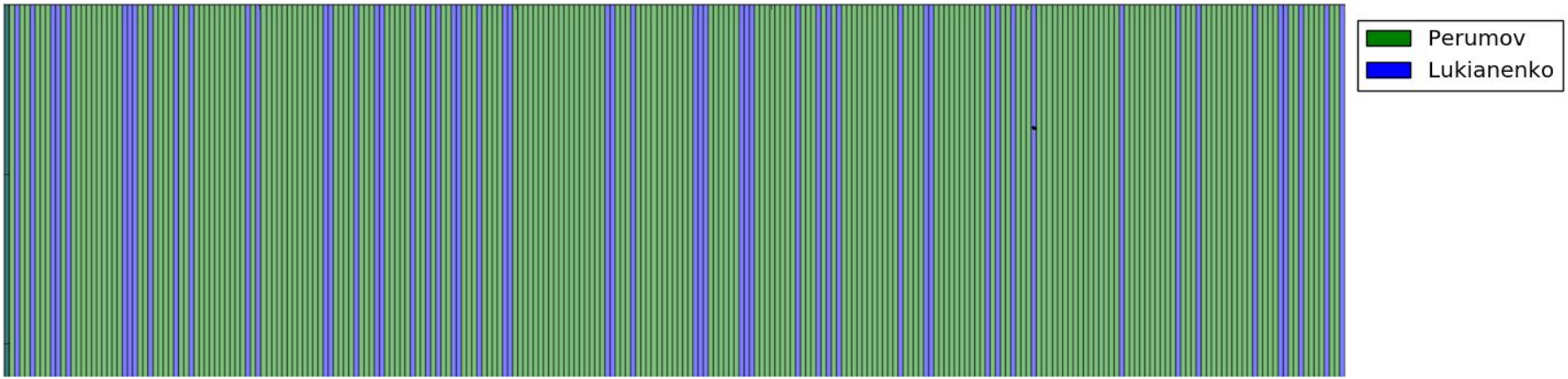
Постановка эксперимента

- Специально собранная библиотека SCI-FI
- Парсинг наиболее популярных электронных книжных форматов
  - PDF, DOC, FB2, RTF, LIT, TXT
- Составлены 8 обучающих корпусов для 4 пар авторов
- Предсказания на текстах в соавторстве

Точность моделей на корпусе



Спектральная картина “Не время для драконов”



Результаты

- Разработан фреймворк для классификации текстов посредством обучения с учителем
- Построена модель, строящая спектральную картину авторства
- Точность предсказания на корпусе SCI-FI достигает 80%
  (по сравнению с 65-70% в методах из статей)

https://github.com/Rauf-Kurbanov/autorship-classification