Санкт-Петербургский государственный университет

Кафедра информационно-аналитических систем

Определение стиля библиографической записи с помощью машинного обучения

Артём Юрьевич Копань, группа 21.Б10

Научный руководитель: ассистент кафедры ИАС Г. А. Чернышев

Санкт-Петербург
2023

Введение

- MAP[1] — сервис для автоматизированной проверка текстов курсовых и ВКР на стилистические ошибки
- Проверка стиля списка литературы на соответствие ГОСТ
- Определение конкретного библиографического стиля по тексту библиографической записи

[1] `https://github.com/Darderion/map`

Постановка задачи

Целью работы является разработка алгоритма для предсказания стиля библиографической записи по тексту записи

Задачи:

- Выявить признаки, по которым различаются библиографические стили, и составить список их основных разновидностей
- Найти существующие алгоритмы, решающие поставленную задачу, описать их достоинства и недостатки
- Найти достаточное количество файлов стилей и библиографий и сформировать датасет
- Разработать и реализовать алгоритм выделения и предобработки значимых признаков в библиографических записях
- Провести экспериментальное исследование и сделать вывод, какой алгоритм машинного обучения лучше справляется с поставленной задачей

Обзор — библиографические стили

- Библиографический стиль — это набор правил, регулирующих внешний вид текста библиографической записи в LaTeX
- Декларирует порядок следования имени автора, названия статьи, издательства и т.д., вид ссылки на источник в тексте
- Пример: **ugost2008s**
  [1] Alander Jarmo, Moghadampour Ghodrat, Ylinen Jari. Comparison of elevator allocation methods // Proceedings of the Second Nordic Workshop on Genetic Algorithms and their Applications (2NWGA) / Ed. by Jarmo T. Alander. — Proceedings of the University of Vaasa, Nro. 13. — Vaasa (Finland) : University of Vaasa, 1996. — 19.–23. . — P. 211–214.

Обзор — существующие решения

- Инструмент[2] от Dominika Tkaczyk (Crossref) решает поставленную задачу с точностью до 94% с использованием линейной регрессии
- Недостатки:
  - Алгоритм выделения признаков не учитывает расстановку пробелов
  - Используется всего 17 библиографических стилей

[2] `https://gitlab.com/crossref/citation_style_classifier`

Создание датасета

- Файлы библиографий (bib) и стилей (bst) скачаны из открытых источников[3, 4], всего 91 стиль, 301 bib файл
- PDF-файлы генерировались автоматически с помощью заранее созданного LaTeX-шаблона и скрипта на Bash
- Перебор всех bib и bst файлов производился Python-скриптом
- Было сгенерировано более 19 тысяч PDF-файлов с 6 млн библиографических записей

[3] Библиографии: `http://dblp.org/`
[4] Стили: `http://ctan.org/`

Выделение признаков (первый способ)

- Библиотека PDFium для парсинга сгенерированных PDF
- В каждом файле выделялись отдельные библиографические записи, а затем для них подсчитывались числовые и категориальные признаки
- Признаки, типы:
  - количество различных знаков пунктуации
  - доля больших букв
  - наличие слов “Abstract”, “Key”, “[s.n.]” и т.д.
  - тип ссылки на запись ([1], [TrD90], [Ski(2022)], [SDD, 99])

  (всего 22 признака)
- Подсчитанные признаки были сохранены в CSV-файле

Выделение признаков (второй способ)

Векторизация

- Конкретные слова заменяются на токены:
  - слово с большой буквы, слово в верхнем регистре, другое слово
  - большая буква, маленькая буква
  - год
  - число
  - пробел
  - знаки пунктуации остаются
- К преобразованному тексту применяется векторизация, т.е. выделение n-грамм с помощью CountVectorizer
- N-грамма — последовательность слов длины N. N-граммы используются в качестве новых признаков

Модели машинного обучения

Для эксперимента были выбраны самые распространённые модели классификаторов

- Линейная классификация — SGDClassifier
- Наивная байесовская классификация — GaussianNB*, MultinomialNB**
- Случайный лес — RandomForestClassifier
- Дерево решений — CatBoostClassifier

Метрики

- Доля правильных ответов алгоритма (accuracy score):

$$\texttt{accuracy}(y, \hat{y}) := \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$

- Точность (precision):

$$P(A, B) := \frac{|A \cap B|}{|B|}$$

- Полнота (recall):

$$R(A, B) := \frac{|A \cap B|}{|A|}$$

- F_1:

$$F_1(A, B) := \frac{2 \times P(A, B) \times R(A, B)}{P(A, B) + R(A, B)}$$

Результаты экспериментального исследования

Таблица: Результаты на первом варианте предобработки признаков

Алгоритм	Accuracy	Precision	Recall	F_1
Линейная классификация	0.41	0.41	0.40	0.37
Наивная байесовская классификация**	0.36	0.42	0.37	0.33
Случайный лес	0.54	0.55	0.52	0.50
Дерево решений	0.62	0.60	0.61	0.60

Таблица: Результаты на втором варианте предобработки признаков

Алгоритм	Accuracy	Precision	Recall	F_1
Линейная классификация	0.66	0.65	0.62	0.61
Наивная байесовская классификация*	0.55	0.58	0.55	0.55
Случайный лес	0.68	0.67	0.65	0.64
Дерево решений	0.84	0.83	0.82	0.82

Результаты

- Составлен список основных разновидностей библиографических стилей LaTeX с примерами
- Проведён обзор существующих решений поставленной задачи, выявлены их достоинства и недостатки
- Сформирован датасет для обучения моделей машинного обучения
- Разработаны и реализованы два различных способа выделения и предобработки признаков
- Проведено экспериментальное исследование, выяснено, что наилучшую точность классификации обеспечивает алгоритм дерева решений, оптимизированный градиентным бустингом

Реализацию и использованные данные можно найти в репозитории[5] GitHub.

[5] https://github.com/ArtyomKopan/Practical_Training_1