СТИЛОМЕТРИЧЕСКИЕ ПРИЗНАКИ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ АВТОРА ТЕКСТА

Труханович И.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Парамонов А.И. – канд. техн. наук

Определение стилометрических признаков в задаче идентификации авторства является одним из основных этапов. Предложены распределенные по группами стилометрические признаки для моделей авторских стилей и средства их извлечения.

На сегодняшний день задача идентификации автора текста является одной из важнейших среди разновидностей обработки естественных языков, поскольку при постоянно растущем количестве анонимных текстов существует необходимость определять автора с заданной точностью.

В случае автоматизированной идентификации автора текста выделяют следующие компоненты решений: признаки авторов, модель представления текста и метод классификации. Одними из лучших методов классификации являются методы машинного обучения [1]. Вдобавок могут быть использованы методы статистического анализа [2].

Одним из первоначальных и концептуально определяющих этапов в задаче идентификации автора текста является выбор признаков, которые могут формировать предполагаемые модели авторских стилей.

Можно выделить пять основных групп признаков: символьные, лексические, синтаксические, семантические и особые. Последние представляют собой признаки, которые предполагают привязку к структурным или языковым особенностям, в то время как остальные являются значительно более универсальными.

В таблице 1 приведены основные представители каждой из групп с указанием инструментов для работы с ними.

Таблица 1 – Признаки авторов

Группа	Название	Описание	Инструменты
Символьные	n-граммы	Частоты символьных n-грамм	Символьный парсер
Символьные	Символьный запас	Множество символов	Символьный парсер
Лексические	n-граммы	Частоты словесных n-грамм	Токенизатор
Лексические	Словарный запас	Множество слов	Токенизатор
Лексические	Ошибки	Орфографические ошибки	Валидатор текста
Синтаксические	Части речи	Частоты частей речи	Разметчик частей речи
Синтаксические	Фразы	Модели построения фраз	Токенизаторы, разметчики частей речи
Синтаксические	Ошибки	Синтаксические ошибки	Валидатор текста
Семантические	Антонимы	Варианты антонимов	Тезаурус
Семантические	Синонимы	Варианты синонимов	Тезаурус
Особые	Структурные	Особенности структуры текстов	Специальные парсеры
Особые	Языковые	Особенности языков	Языковые справочники

Одной из основных особенностей является то, что одни признаки могут быть эффективны как в одиночку, так и в совокупности, в то время как другие могут служить, как правило, лишь в качестве дополнительных. Как правило, частоты символьных и лексических групп показывают высокую эффективность на малых масштабах без использования дополнительных признаков.

Семантические признаки являются менее независимыми, но в случае большого количества авторов могут работать гораздо лучше, а тезаурус позволяет применять более гибкие подходы в зависимости от задачи [3].

Список использованных источников:

1. Труханович, И. А. Обзор решений задачи идентификации автора текста / И. А. Труханович, В. С. Кунцевич // Компьютерные системы и сети: 57-я научная конференция аспирантов, магистрантов и студентов, Минск, 19-23 апреля 2021 г. : сборник тезисов докладов / Белорусский государственный университет информатики и радиоэлектроники. – Минск, 2021. – С. 61.

2. Парамонов, А. И. Методы анализа цифрового текста для идентификации его автора / А. И. Парамонов, И. А. Труханович // Веб-программирование и интернет-технологии WebConf2021 : материалы 5-й международной научно-практической конференции, Минск, 18-21 мая 2021 г. / Белорусский государственный университет ; редкол.: И. М. Галкин [и др.]. – Минск, 2021. – С. 118–119.

3. Top Thesaurus APIs [Электронный ресурс]. – Режим доступа: https://rapidapi.com/collection/thesaurus-apis. – Дата доступа: 02.04.2022.