Название: Определение оптимальных параметров разбиения текста на сегменты при решении задачи обнаружения внутреннего плагиата.

Задача: Рассматривается документ d — последовательность символов $c_1, \ldots, c_{L_d}$. В документе возможно наличие чужеродных блоков. Требуется сопоставить каждому символу c_l документа метку класса

$$t_l = \begin{cases} 1, & \text{если символ принадлежит чужеродному блоку,} \\ 0, & \text{иначе,} \end{cases} \qquad l = 1, \ldots, L_d.$$

Большинство решений [1–3], предложенных для данной задачи, построено по схеме, включающей разбиение документа на сегменты s_i, $i = 1, \ldots, m$ (абзацы, предложения, блоки слов или символов), профилирование сегментов — выделение признаков $\mathbf{x}_i \in \mathbf{R}^n$, и выделение аномальных сегментов. На этом этапе построенное признаковое описание $\mathbf{x}_i$ используется для сравнения сегментов s_i и выделения сегментов, принадлежащих чужеродным блокам. Здесь используются методы классификации либо обнаружения выбросов.

Качество признаков зависит от способа разбиения текста на сегменты.
Гипотеза: для каждого признака существует оптимальный способ разбиения текста на сегменты.
Задача: ввести (или обоснованно выбрать одну из предложенных ранее [4, 5]) меру качества признака, протестировать применимость различных пар <признак, способ разбиения текста>.

Базовой алгоритм: Способ разбиения выбирается согласно экспертным соображениям. Существуют рекомендации по выбору размера сегмента, основанные на исследованиях стабильности. Пример: [4]. Для определения оптимальной длины сегмента, рассматривается набор однородных текстов. Разбиение считается успешным, если для всех сегментов значение признака совпадает со значением, полученным для всего документа. Оптимальной длиной считается та длина сегмента, которой соответствует максимальное количество успешных разбиений.

Решение: * Рассмотреть набор пар <признак, способ разбиения текста> для наиболее успешных признаков. Для фиксированной пары исследо-

вать следующие свойства: качество F_1 решения задачи для данного признака, signal to noise ration, odds ratio, Husrt index etc.
* В качестве этапа пост-обработки в дополнение к стандартным вариантам связных блоков стоит рассмотреть вариант со случайным сэмплированием из работы [6]

Список литературы

[1] Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style, Gabriel Oberreuter, Juan D. Velasquez, 2013

[2] Intrinsic Plagiarism Detection Using Character n-gram Profiles Efstathios Stamatatos, 2009.

[3] Benno Stein, Sven Meyer zu Eissen. Intrinsic Plagiarism Analysis with Meta Learning, 2007.

[4] Simon Suchomel and Michal Brandejs. Determining Window Size from Plagiarism Corpus for Stylometric Features, 2015.

[5] Moshe Koppel, Navot Akiva, Ido Dagan. Feature Instability as a Criterion for Selecting Potential Style Markers, 2006.

[6] Maciej Eder. Does size matter? Authorship attribution, small samples, big problem Digital Scholarship Humanities (2015) 30 (2): 167-182