Разбор статьи "Methods for intrinsic plagiarism detection and author diarization" Mikhail Kuznetsov, Anastasia Motrenko, Rita Kuznetsova, and Vadim Strijov

Говоров Иван, МФТИ

**Постановка задачи**

Метод обнаружения плагиата, основанный на структурировании функции стиля автора из особенностей текстовых предложений и обнаружения выбросов.

**Данные**

Было проведено несколько экспериментов на данных из коллекции PAN-2011.

Было взято 4753 статьи, которые в свою очередь были разделены на 10 групп.

**Предложенный метод решения**

Предлагаемый метод обнаружения плагиата разбивает текстовые документы на предложения. Векторизирует их и обучая классификационную модель находит отклонения в классификаторах. Дополнительно сегментируя данные в набор кластеров соответствующих каждому автору, для приспособления метода к проблеме диалектики автора.

**Актуальность / обоснование метода**

Позволяет выявлять плагиат в статьях где четко выражен автор которому принадлежит не менее 70 процентов статьи с вставками фрагмето быть может от других авторов. Что на сегодняшний день, с публикацией огромного числа статей, имеет критическое значение.

**С чем сравнивают**

Сравнение происходит по мерам предложенным на соревнования PAN.

**Результаты**

Данный метод был реализован на конкурсе PAN-2016, модель смогла достичь f1 меры 0,2 для внутренней проблемы плагиата. bcubed-f меры 0,54 для статей с известным числом авторов и bcubed-f меры 0,5 для статей с неизвестным числом авторов.