КЛАСТЕРИЗАЦИЯ ТЕКСТОВ НА ОСНОВЕ МОДЕЛИ СЛУЧАЙНЫХ БЛУЖДАНИЙ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ ПАРАМЕТРОВ

Крамаренко А.А.

Уральский федеральный университет имени первого Президента России

Б.Н. Ельцина, г. Екатеринбург, Россия

E-mail: lonelywolf1333@gmail.com

TEXTS CLUSTERING BASED ON THE RANDOM WALK MODEL. LAWS OF PARAMETERS DISTRIBUTION

Kramarenko A.A.

Ural Federal University, Yekaterinburg, Russia

Annotation. Authors got parameter values for 120 texts analyzed using the random walk model and described them by the theoretic gamma distribution. Also they estimated statistical significance of texts clustering.

Всё больше растёт интерес к машинной обработке языка и автоматизированному анализу текстов. Многие задачи в этой области связаны с классификацией текстов на основе каких-либо критериев. Это может быть полезно для анализа большого количества текстов с целью выявления каких-либо особенностей, идентификации авторства текста, определения спама, в информационном поиске и так далее.

В статьях [1,2] подробно описан разработанный авторами подход для кластеризации текстов по определённым критериям, основанный на математической модели случайных блужданий.

В рамках этого подхода анализируемые тексты характеризуются коэффициентом диффузии (параметр D) и относительной поправкой к закону Эйнштейна (параметр RC). Первый параметр является интегральной характеристикой частот появления букв и отражает некие структурные особенности текста. Второй представляет собой меру неоднородности текста, изменения свойств текста от его начала к концу.

Было исследовано 120 текстов, разделённых на шесть групп в соответствии со стилем написания текста или его тематикой. Полученные значения параметров D и RC были проанализированы и описаны гамма-распределением с параметрами k и θ , подобранными для каждой группы. По характеру распределений и их соотношениям друг с другом можно проследить закономерности как по текстам внутри стиля, так и внутри текста в соответствии со стилем.

Выяснилось, что статистически значимо ($\alpha = 0,1$) по обоим параметрам отличаются друг от друга почти все распределения, за исключением распределений прозы и публицистики. Необходимо отметить, что распределение параметров для текстов художественной прозы почти полностью перекрыто распределениями других групп текстов.

Поведение распределений выборок текстов из подкорпусов по параметру RC не совпадает с их поведением по параметру D , то есть эти параметры не эквивалентны и кластеризация текстов по параметру RC дополняет кластеризацию текстов по параметру D .

По выделенным параметрам была проведена кластеризация контрольных выборок текстов, на основе которых в дальнейшем возможна классификация текстов по стилям или каким-либо другим критериям. Также разработанная нами модель может служить основой комплекса программ для кластеризации текстов.

1. Крамаренко А.А., Филимонов В.В. и др., Информация: передача, обработка, восприятие, с. 138-146 (2017).
2. Kramarenko A.A., Nekrasov K.A. et al., AIP Conference Processing, 1886, p. 020072 (2017).