Ранжирование текстов литературных произведений

Рысьмятова Анастасия Александровна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: rysmyatova@gmail.com

Автоматическая обработка текстов становится все более востребована в связи с постоянно растущим объемом информации в Интернете и потребностью в ней ориентироваться. Ранжирование текстов — важная задача автоматической обработки текстов, исследованиями в которой активно занимаются все поисковые системы. Наиболее популярные поисковые системы используют методы машинного обучения для решения данной задачи.

Задачу ранжирования текстов можно формализовать следующим образом [1]: X — множество объектов; $X^\ell = \{x_1, ..., x_\ell\}-$ обучающая выборка; $i \prec j$ — правильный порядок на парах $(i, j) \in \{1, ..., \ell\}^2$; необходимо построить ранжирующую функцию $a : X \to \mathbb{R}$ такую, что $i \prec j \Rightarrow a(x_i) < a(x_j)$.

В работе исследованы основные подходы к ранжированию текстов на примере задачи ранжирования литературных произведений. Для этого с сайта [4] выбраны тексты стихотворений различных русских поэтов и, используя методы машинного обучения, построен алгоритм способный ранжировать тексты одного автора в порядке возраста, в котором он написал произведения.

Для решения данной задачи ранжирования был применен попарный подход: был построен бинарный классификатор, принимающий на вход пары стихотворений одного и того же автора, и определяющий, какое стихотворение было написано раньше. Качество классификации измерялось по метрике Accuracy (доля верно классифицированных объектов).

В работе приведены результаты решения задачи ранжирования литературных произведений с использованием как нейросетевых методов машинного обучения, получивших особую популярность в последнее время [2–3], так и традиционных методов автоматической обработки текстов.

В работе показана возможность предположить возраст автора, в котором было написано стихотворение, если известна информация о других произведениях данного автора.

Литература

1. Воронцов К. В. Курс лекций по машинному обучению, 2015.
2. Kim Y. Convolutional neural networks for sentence classification// In IEMNLP, 2014, P. 1746–1751.
3. Zhang X, Zhao J, Yann LeCun. Character-level convolutional networks for text classification // In Advances in Neural Information Processing Systems, 2015, P. 649–657.
4. Страница сайта «World-Art»: `http://www.world-art.ru/lyric/`