План статьи журнала «Информационные технологии» №6, 2017

1. Обработка кодов УДК
2. Искусственная нейронная сеть
3. Генерация кода УДК

Тезисы статьи журнала «Информационные технологии» №6, 2017

А. Ю. Романов, канд. техн. наук, ст. преп.,
К. Е. Ломотин, студент.
Е. С. Козлова, студентка.
Национальный исследовательский университет «Высшая школа экономики»

Применение методов машинного обучения для решения задачи автоматической рубрикации статей по УДК

На современном этапе задача обработки текстов на естественных языках является одной из насущных. Она напрямую связана с растущим количеством текстовой информации в сети Интернет. Научные статьи, книги, руководства, журналы, справочники – все это многообразие текстов в большинстве в своем представлено в цифровом виде.

В данном случае в качестве обучающего набора данных выступают тексты научных статей, а в качестве ответа эксперта – код УДК, присвоенный статье автором или модераторами ресурса. Центральной частью УДК являются основные таблицы, охватывающие всю совокупность знаний и построенные по иерархическому принципу деления от общего к частному с использованием цифрового десятичного кода.

Нейронные сети прямого распространения в последнее время обрели большую популярность и их успешно применяют для классификации различных объектов, в том числе и текстов. Анализ подобных поверхностей может являться обоснованием для автоматического подбора гиперпараметров при построении системы классификаторов. Все нейронные сети в ходе экспериментов обучались по градиентному методу «Adam» с адаптивной скоростью обучения, разработанному командой ученых из Амстердама (Нидерланды) и Торонто (Канада).

Цель данного исследования – разработать систему, которая сможет с высокой точностью генерировать коды УДК для научных статей, основываясь на принципах машинного обучения.

В процессе исследования были разработаны графовые алгоритмы генерации кода УДК, комбинация которых отличается гибкостью и высоким потенциалом для улучшения. Одним из существенных результатов работы стал прототип программного модуля, работающий по принципу логистической регрессии и способный генерировать первую цифру кода УДК.

Стр. 418-423