Университет ИТМО, факультет инфокоммуникационных технологий

Отчетная работа по «Информатике»: аннотация к статье

Выполнил(а) Долматов Д.А. (ФИО студента), № группы K3221, дата 20.11.2021, оценка _____ (не заполнять)

Название статьи/главы книги: Классификация медиатекстов с использованием машинного обучения		
ФИО автора статьи: Л.В. Мотовских	**Дата публикации:** 2020	**Размер статьи** 7 стр.
Прямая полная ссылка на источник и сокращенная ссылка: https://cyberleninka.ru/article/n/klassifikatsiya-mediatekstov-s-ispolzovaniem-mashinnogo-obucheniya/viewer https://clck.ru/Yw8Bc		
Тэги, ключевые слова или словосочетания: классификация текстов; TF-IDF; «случайный лес»; метод опорных векторов; электронные СМИ.		
Перечень фактов, упомянутых в статье: Методы машинного обучения намного эффективнее оптимизируют работу с большими данными по сравнению с классическими способами алгоритмирования способов получения и обработки информации. Одной из сфер применения алгоритмов нейронных сетей применяется в классифицировании информации в СМИ. В статье описывается два алгоритма, а источником медиатекстов являются внутренние СМИ РФ, которые имеют подразделение на: общество, экономику, политику, спорт и культуру. Сначала собирается выборка, при этом отбрасываются 5% как самых длинных статей, так и 5% самых коротких по количеству символов. Итоговая выборка подразделяется на две в соотношении 85:15 на обучающую и проверочную. Текстовый формат информации представляется в терм-документной таблице - матрицы TF-IDF, после чего применяется алгоритм «Random forest», который случайно распределяет элементы подвыборки каждой матрицы (дереву) и присваивает им случайный набор элементов с количеством деревьев, узлов, вершин, глубин. Точность таких результатов равна 86%. Второй алгоритм «Метод опорных векторов» разделяет подвыборку на несколько гиперплоскостей, которые находятся на максимально возможном расстоянии друг от друга. Предсказание осуществляется через положение элемента относительно оптимальной гиперплоскости. Для присваивания первичного набора значений используется метод случайного поиска, а затем поиск по сетке параметров. Результаты классификации выборки медиатекста оказались равны 88%, что оправдало ожидания исследователей.		
Позитивные следствия и/или достоинства описанной в статье технологии 1) Поиск по сетке параметров явно превосходит быстрый поиск по точности. 2) Точное предсказывание модели классификатора для основных рубрик СМИ. 3) Возможность классификации большего количества неразмеченных текстов других СМИ.		
Негативные следствия и/или недостатки описанной в статье технологии 1) Наибольшая погрешность классификатора между категориями «Общество» и «Спорт». 2) Необходимость больших вычислительных мощностей для построения групп TF-IDF. 3) Отсутствие стратегического значения данной технологии (кроме сортировки писем СПАМа).		
Ваши замечания, пожелания преподавателю или анекдот о программистах Such a great day today!		