ЗАДАЧА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

С.А. Никитина, к.ф.-м.н., доц.
Челябинский государственный университет
e-mail: *nikitina@csu.ru*
С.С. Попов, магистрант
Челябинский государственный университет
e-mail: *sergei.popov174@gmail.com*

Решается задача автоматической классификации текстов. Определяется тематическая принадлежность текста на русском языке. Для проведения классификации применены методы машинного обучения. Методы протестированы на разных наборах параметров, выявлены наиболее существенныеиз них. Написановеб-приложение, которое позволяет выполнять анализ текста и определять его тематическую принадлежность.

Ключевые слова: автоматическая классификация текстов, определение тематической принадлежности текста, веб-приложение.

1. Введение

Классификация текстов на сегодняшний день является одной из актуальных задач, поскольку к ней сводится ряд других задач: определение автора текста,тематической принадлежности текстов, эмоциональной окраски высказываний и др.

Один из подходов к классификации основывается на методах машинного обучения [3]. При классификации документов категории (классы) определены заранее. Процесс классификации заключается в следующем. Задается набор примеров, который называют обучающей выборкой. Обучающую выборку используют для обучения классификатора и определения значения параметров, при которых классификатор выдает лучший результат. Затем в системе вырабатываются правила, с помощью которых происходит разделение множества на заданные классы. Качество классификации проверяется тестовой выборкой.

Многие из существующих систем классификации текстов ориентированы на англоязычные коллекции текстов [6]. В данной статье решается задача автоматической классификации текстов на русском языке по темам с использованием методов машинного обучения и методов естественной обработки языка. Разработано веб-приложение, позволяющего производить данный анализ автоматически.

2. Методика проведения исследований

Классификация состоит из нескольких последовательных этапов: предобработка и индексация, построение и обучение классификатора, оценка качества классификации [1].

При предобработке текста можно встретить достаточно большое количество корректных словоформ со схожими значениями, написания которых могут отличаются различными изменяемыми частями слова (например, приставками, суффиксами, окончаниями и т.д.). Это усложнит дальнейшую обработку текста, а также создание словарей. Лемматизация позволяет привести слово к его сло-

варной форме − лемме. Предварительная обработка текста значительно сокращает размерность пространства.

Индексация документов предполагает построение некоторой числовой модели текста, которая переводит текст в удобное для дальнейшей обработки представление. Один из распространенных методов индексации – Word2vec, в котором каждое слово представляется в виде вектора, содержащего информацию о сопутствующих словах.

Существуют несколько способов определения веса признаков документа. Наиболее распространенный – вычисление функции TF-IDF. Его основная идея состоит в том, чтобы больший вес получали слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

В данной работе использовались два метода обработки текста: лемматизация и векторизация TF-IDF.

На сегодняшний день разработано большое количество методов машинного обучения. Для решения поставленной задачи были выбраны следующие: метод опорных векторов (SVM), метод ближайших соседей (KNN), метод на основе логистической регрессии (LogReg).

Суть метода опорных векторов [1, 2] заключается в построении гиперплоскости, разделяющей имеющиеся объекты наилучшим образом. Причем, в алгоритме предполагается, что чем больше расстояние между разделяющей гиперплоскостью и объектами классов, тем меньше ошибок будет допущено в работе алгоритма. Зачастую структура данных бывает неизвестна и очень редко удается построить разделяющую гиперплоскость. В таком случае необходимо перейти от исходного пространства признаков документов к новому, в котором обучающая выборка окажется линейно разделимой. Для этого каждое скалярное произведение заменяют на некоторую функцию, отвечающую определенным требованиям. Вообще, задача построения наилучшей разделяющей гиперплоскости сводится к задаче квадратичного программирования.

В основе метода ближайших соседей [1, 5] лежит оценивание расстояний до объектов. Для повышения надежности и точности классификации, объекты необходимо относить к такому классу, к которому принадлежат k его ближайших соседей обучающей выборки.

Логистическая регрессия [1, 4] – еще один метод построения классификатора, с помощью которого оцениваются апостериорные вероятности принадлежности объектов классам. Задаются зависимая переменная, принимающая одно из двух значений и множество независимых переменных, на основе значений которых вычисляется вероятность принятия значения зависимой переменной. При классификации документов в качестве зависимой переменной выступает класс (тематика), а в качестве независимых переменных – множество документов.

3. Обсуждение полученных результатов

В качестве инструментов для решения задачи был выбран язык программирования Python и библиотеки к нему sklearn, pandas, pymorphy2.

Была проведена лемматизация с помощью библиотеки pymorphy2, а также tf-idf векторизация с помощью модуля TfidfVectorizer из библиотеки sklearn. Далее использовались модули, LinearSVC, LogisticRegression, KneighborsClassifier из библиотеки sklearn, с разными параметрами. Замерялась скорость обучения и предсказания, а также точность предсказания (таблица 1-4).

Таблица 1– Результаты работы методов с параметрами из «коробки»

Метод	Время обучения на 20000 объектов, сек	Время предсказания на 5000. объектов, сек	Точность, %
Метод опорных векторов (SVM)	16.2	0.258	89.94
Логистическая регрессия (LogReg)	59.3	0.258	80.45
Метод ближайших соседей (KNN)	0.85	231	77.16

Таблица 2 – Результат SVM с перебором параметра *C*

Значение *C*	Точность
5	90
10	90
7.5	89.91
3	89.93
6.5	89.87
0.5	89.5
0.1	85.92

Таблица 3 – Результат LogReg с перебором параметра *C*

Значение *C*	Точность
3	85.73
5	86.97
10	88.17
15	88.52
0.5	75.4
0.1	58.89

Таблица 4 – Результат KNN с перебором параметра *k*

5	77.16
10	76.30
15	75.59
30	73.45
35	72.73
50	71.59
100	68.53

После проведения численных экспериментов в качестве наилучшей модели была выбрана KNN. На её основе так же проводился перебор параметров векторизации tf-idf (таблица 5).

Таблица 5 – Результат KNN с разными параметрами tf-idf

max_features	ngram_range	Точность SVM
100000	Нет	77.77
50000	Нет	76.62
1000000	2	78.44
Нет	2	82.75
700000	2	80.71
100000	3	83.30
1000000	3	85.14

Было разработано демонстрационное веб-приложение на языке Python, с использованием микрофреймворка Flask. Данное приложение позволяет определять для произвольного текста на русском языке его тематику.

4. Заключение

Рассмотрена задача определения тематической принадлежности текста на русском языке.В рамках решения задачи выполнен обзор алгоритмов классификации и алгоритмов приведения текстов к векторному виду.

Для проведения классификации были реализованы:метод опорных векторов (SVM), метод ближайших соседей (KNN), метод на основе логистической регрессии. Для реализации веб-приложения была выбран SVM. Алгоритм классификации протестирован на разных наборах характеристик и параметров, выявлены наиболее значимые характеристики,определенылучшие параметры.

Реализованное в ходе работы приложение показало свою работоспособность при проведении экспериментов. Данное приложение может быть модернизировано в дальнейшем путём расширения тренировочной выборки, применения других способов векторизации текстов.

ЛИТЕРАТУРА

1. Батура Т.В. Методы автоматической классификации текстов // Программные продукты и системы. –2017. – Т. 30. – № 1. С. 85–99.
2. Воронцов, К. В. Лекции по методу опорных векторов [Электронный ресурс] // URL: https://yourcmc.ru/wiki/images/6/6a/Vorontsov-SVM-draft.pdf.
3. Барсегян А.А., Куприянов М.С., Холод И.И. и др. Анализ данных и процессов. –СПб.: БХВ-Петербург.–2009. – 512 с.
4. Логистическая регрессия: [Электронный ресурс]. // [сайт] URL: http://www.machinelearning.ru/wiki/index.php?title=Логистическая_регрессия.
5. Метод ближайших соседей: [Электронный ресурс]. [сайт] // URL: http://www.machinelearning.ru/wiki/index.php?title=KNN.
6. Korobov M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts, P 320-332 (2015).