электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.93'1

Автоматическая классификация текстовых документов на русском и английском языках с помощью методов машинного обучения

Н.Д. Лыфенко

Студент, «Кафедра математики, логики и интеллектуальных систем в гуманитарной сфере» РГГУ, г. Москва, Россия

Научный руководитель: А.В. Бобков, к.т.н., доцент кафедры «Системы автоматического управления» МГТУ им. Н.Э. Баумана, г. Москва, Россия

РГГУ
LyfenkoNick@yandex.ru

В связи с увеличением широты информационных потоков в различных сферах профессиональной деятельности возникает потребность в автоматической классификации разнородных документов, т.е. соотнесении документа одному или нескольким классам, исходя из его содержания. В статье предлагается один из возможных способов решения этой задачи.

Существуют различные методы классификации текстов – деревья принятия решений, нейронные сети, линейные классификаторы и другие [1]. В качестве подхода для классификации документов в соответствии с его темой была выбрана байесовская модель, т.к. в первом приближении дает неплохие результаты и представляется наиболее простой для понимания и программирования.

В данной статье предлагается решение задачи автоматической классификации множества документов по заданному классу – теме или стилю речи для русского и английского языков с применением методов машинного обучения. Выборка документов по конкретному классу состоит из отобранных экспертом текстов.

Каждому тексту ставится в соответствие вектор в n-мерном пространстве, его координатой является частота встречаемости лексемы. Для сокращения размерности пространства и надежности использования метода соответствующие словоформы объединяются в одну лексему. При этом было установлено, что не все слова языка являются важными для данного класса и могут характеризовать его. Поэтому для решения возникшей проблемы используется список стоп-слов русского и английского языков, в

который были включены служебные части речи, некоторые глаголы, наречия, местоимения, которые не являются специфическими для каждого класса. Особенности русского и английского языков привили к тому, что для русского языка используется грамматический словарь русского языка Зализняк А.А. [2], для английского языка – основа слова. Грамматическая омонимия для обоих языков не учитывается.

Непосредственно классификация документов происходит следующим образом. Строится гистограмма на основе векторов, характеризующих каждый класс, и сравнивается имеющийся вектор, представляющий текст для классификации. В результате получаем вероятность принадлежности текста к определенному классу.

В процессе классификации стало ясно, что необходим модуль по настройке классификатора, т.к. каждый новый добавленный текст в имеющуюся тему или стиль вносит изменения: увеличивается частотность лексем характерных для данной темы (стиля), если этот текст принадлежит данному классу, иначе увеличивается частота встречаемости незначимых лексем. Поэтому был реализован механизм по добавлению текста к множеству документов и его удалению.

Для решения этой задачи использовалась модель ядра и периферии, которая состоит в следующем: ядро включает вектора, находящиеся на минимальном расстоянии от центрального вектора, т.е. с наибольшей вероятностью совпадения с выбранной темой (стилем). Периферия, напротив, состоит из векторов, наиболее удаленных от центра. Появление нового вектора приводит к перерасчету центра.

В ходе работы программы, реализующей данный механизм, соотношение количества текстов, находящихся в ядре к количеству текстов из периферии, определяется аппроксимирующей функцией. Поэтому при добавлении нового текста (вектора) необходимо понять, куда он попадет. Вектор, оказавшись в ядре, инициирует пересчет периферии, и наиболее удаленный от центра вектор попадет в периферию. Аналогичная ситуация происходит при добавлении вектора (нового документа) в периферию.

Для проверки работоспособности программы в ходе данного исследования было проведено два эксперимента. При этом тексты для классификации были представлены на русском языке.

Целью первого эксперимента была классификация текстовых документов по следующим темам: *спорт, культура, политика, наука.* Эксперимент состоял в следующем. В качестве обучающей выборки экспертом было отобрано 40 документов, объемом одна страница печатного текста каждый, соответственно, десять документов для каждой из заявленных тем. В качестве тестовой выборки было предложено еще 40 документов с такими же параметрами. В ходе эксперимента программа допустила ряд

ошибок, под которыми понимается неверное отнесение документа к теме на основе знаний эксперта. Результаты эксперимента представлены в таблице 1.

Таблица 1

Результаты эксперимента по классификации текстового документа по темам

Название темы	Культура	Спорт	Политика	Наука
Ошибка	40%	0%	0%	30%

Довольно большие ошибки связаны с маленьким объемом обучающей выборкой и ее спецификой. Большинство текстов были взяты с новостных сайтов (*lenta.ru, rbc.ru*). Высокие ошибки (30%-40%), возможно, связаны с качеством отобранных документов.

Целью второго эксперимента была классификация текстового документа по стилям речи: *научный, художественный, публицистический, официально-деловой, разговорный*. Объем обучающей выборки равен 100 документам. Для каждого стиля было отобрано 20 документов объемом одна страница печатного текста каждый. Объем тестовой выборки составил 100 документов, 20 для каждого стиля. Результаты работы представлены в таблице 2.

Таблица 2

Результаты эксперимента по классификации текстового документа по стилям

Стиль	Научный	Художественный	Публицистический	Официально-деловой	Разговорный
Ошибка	0%	0%	0%	0%	40%

Тексты научного стиля составили тезисы научных конференций, доклады, диссертации, рецензии. Для художественного стиля были отобраны произведения Л.Н. Толстого, А.П.Чехова и рассказы участников конкурса сетевой литературы «Тенета-98». Публицистический стиль представлен данными с новостных сайтов и электронных версий печатных изданий (*lenta.ru, mail.ru, kp.ru*). Для официально-делового стиля были выбраны законы РФ, взятые с официального сайта «Консультант Плюс».

Программа относила только тексты разговорного стиля неправильно. Возможно, это связано со спецификой самих документов, т.к. основную часть составили переписки из чатов.

Таким образом, можно сделать вывод, что представленный подход для классификации текстовых документов даже на небольших данных дает неплохие результаты. В дальнейшем можно стремиться к уменьшению количества ошибок, протестировав программу на большем объеме качественно составленных экспертом

текстов, и понять, на сколько выбранный подход является эффективным. Возможно, в дальнейшем следует перейти от байесовского подхода к более усложненным методам классификации текстовых документов (например, метод опорных векторов) и от представления текстового документа в виде вектора частот встречаемости лексем перейти к более точному, например, *TF-IDF*.

В ходе данного исследования была выполнена поставленная задача, было реализовано приложение по классификации текстовых документов по заданным темам и стилям речи, приложение по настройке классификатора.

Литература

1. Fabrizio S. Machine Learning in Automated Text Categorization // ACM Computing Surveys, 2002.№34(1), pp 1-47.
2. Зализняк А.А. Грамматический словарь русского языка. М.: Русский язык, 1980. 795 с.
3. Введение в информационный поиск /К. Д. Маннинг [и др.] М.: Вильямс, 2011. 528 с.