Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 681.3.016(075)

Веремейчик
Мария Ивановна

Интеллектуализация процесса обработки деловых документов в веб-приложениях

АВТОРЕФЕРАТ
На соискание степени магистра технических наук

по специальности 1-31 80 10 «Теоретические основы информатики»

_____

Научный руководитель

Колб Д.Г.
кандидат техн.наук, доцент
кафедры ИИТ

_____

Минск 2019

ВВЕДЕНИЕ

В мире в последние несколько лет существенно возросло количество информации. Информация накапливается каждый день и становится неотъемлемым ресурсом общества.

В связи с этим появляется проблема структуризации информации. Для того, чтобы в условиях огромных накопленных массивов данных отыскать нужную информацию, нужны методы поиска необходимой информации. Такие задачи относятся к задачам классификации. Классификация текста применяется в решении многих практических задач: поиске документов, навигации в больших информационных ресурсах, фильтрации спама, подборе контекстной рекламы, составлении интернет-каталогов и других. Так как на сегодняшний день большинство документации размещено на веб-ресурсах, нужны алгоритмы или полноценные системы, выполняющие эффективный поиск по выбранным документам.

В этой связи становится крайне актуальной задача интеллектуализации процесса обработки деловых документов в веб-приложениях. Сложность разработки состоит в том, что нужно прежде всего выбрать методы, наиболее подходящие к данной задаче. Методов машинного обучения достаточно много и теоретически любой из них может применяться в решении задачи классификации текста, важно выбрать тот, который даст наилучшие результаты.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования

Классификация текста является актуальной и интересной задачей в условиях постоянно растущего объема информации. Принадлежность к тому или иному классу может определяться общей тематикой текста, употреблением определенных понятий, другими условиями. Т.к. количество документации на любом предприятии велико, а способов ее эффективной обработки нет, задача обработки деловых документов является необходимой в условиях роста документов.

Цель и задачи исследования

Цель диссертации состоит в том, чтобы проанализировать существующие подходы к решению задачи обработки деловых документов в веб-приложениях, выбрать и обосновать наиболее подходящие методы для решения задачи классификации с целью разработки в дальнейшем полноценной системы, осуществляемой поиск по массивам деловой документации на предприятии.

Для осуществления цели необходимо решить следующие задачи:

1. Осуществить разметку документа для последующей обработки с помощью методов машинного обучения. Произвести обзор видов признаков, выделить классы.

2. Проанализировать алгоритмы, используемые для извлечения признаков, выбрать наиболее подходящие.

3. Провести вычислительный эксперимент по классификации документов, проанализировать полученные результаты.

Объектом исследования являются методы и алгоритмы машинного обучения, способные решить задачу классификации.

Предметом работы выступают документы.

Теоретическая и методологическая основа исследования

В основу диссертации легли эксперименты и теоретические знания преподавателей Факультета компьютерных наук НИУ ВШЭ и Школы Анализа Данных Яндекса, также их лекции и научные статьи.

Для получения теоретических результатов исследования применялись научные статьи и описания создателей курса “Введение в машинное обучение” от преподавателей Школы Анализа Данных Яндекса.

Имитационные результаты были получены с применением языка программирования Python, библиотек NumPy, SciPy, Pandas, Scikit-Learn, которые включают в себя различные методы и алгоритмы машинного обучения.

Информационная база исследования для анализа сформирована на основе выбранных документов. Были использованы программы поступления в магистратуру различных высших учебных заведений.

Научная новизна диссертационной работы заключается в построении оригинального решения поиска структурных частей в документах, используя методы машинного обучения с возможностью расширения данного решения до полноценной поисковой системы.

Основные положения, выносимые на защиту

1. Выделены классы и признаки документов для решения задачи классификации.

2. Выделен класс алгоритмов для поиска структурных частей документа.

3. Проведен эксперимент, показывающий точность выбранного алгоритма на тестовых данных.

Теоретическая значимость диссертации заключается в том, что в ней проанализированы наиболее популярные используемые методы машинного обучения для решения задач классификации и сделан обоснованный вывод, которые из них работают наиболее эффективно.

Практическая значимость диссертации состоит в том, что на основе

предложенного анализа можно разрабатывать полноценную систему по структуризации деловой документации.

Структура и объем работы. Структура диссертационной работы обусловлена целью, задачами и логикой исследования. Работа состоит из введения, трёх глав и заключения, библиографического списка. Общий объем диссертации – 60 страницы. Работа содержит 12 таблиц, 17 рисунков. Библиографический список включает 14 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** рассмотрено современное состояние проблемы структуризации больших объемов документации, определены основные направления исследований, а также дается обоснование актуальности темы диссертационной работы.

В **общей характеристике работы** сформулированы ее цель и задачи, показана актуальность проводимых исследований, даны сведения об объекте исследования.

В первой главе рассматриваются этапы и методы поиска информации в веб-приложениях, происходит структуризация деловых документов. На основе структуризации документов выделяются классы и признаки, описывающие классы, проведен сравнительный анализ признаков, сделаны выводы о их встречаемости, пересекаемости и весе в каждом документе.

Во второй главе рассмотрен процесс интеллектуализации процесса обработки деловой документации. Проанализированы наиболее встречающиеся методы машинного обучения, позволяющие решить задачу классификации текста. Проведена оценка таких методов как решающие деревья, к ближайших соседей, опорных векторов, Байеса, нейронные сети, сверточные нейронные сети. Дан краткий анализ вышеописанных методов.

В **третьей главе** представлены результаты эксперимента. Представлены инструменты, с помощью которых осуществлялась программная реализация, проведен анализ полученных данных.

ЗАКЛЮЧЕНИЕ

В результате магистерской работы обобщены и систематизированы подходы, решающие задачу классификации текста. Были проанализированы наиболее популярные методы классификации деловой документации, в том числе и методы классификации текстов с использованием нейронных сетей.

Также в данной работе было рассмотрено два основных метода использования сверточных нейронных сетей для задачи классификации текста: посимвольный подход и подход с использованием кодирования слов. Показано, что данная нейронная сеть справляется с задачей классификации текстов лучше, чем все остальные рассмотренные методы. Исследовано влияние на качество классификации некоторых алгоритмов предобработки текста.

Принимая во внимание все вышесказанное можно утверждать, что поставленные в начале работы цели были достигнуты и все поставленные задачи выполнены.

В будущем разработка данного алгоритма может быть использована в полноценной системе поиска структурных частей документа для обработки больших объемов документации на предприятиях.