Проанализировав полученные результаты от проведенного социального опроса, приходим к выводу, что голосовые ассистенты смогут оказать большую помощь в работе, но имея при этом свои недостатки, с которыми можно бороться и совершенствоваться дальше.

Если говорить о современном производстве, подобные «помощники» уже являются полноправными сотрудниками, ускоряющими работу и повышающими эффективность производства. Такой поддержкой и приходится «ATHENA». На данный момент она существует как дополнение к управлению станком, которое предоставляет пользователям работать с блоком управления.

Кардинально облегчить адаптацию наладчиков и операторов. Система состоит из гарнитуры с функциями шумоподавления и программного обеспечения, работающими на ноутбуке. Оператор использует микрофон гарнитуры, чтобы давать команды и задавать вопросы - практически любую команду, которую можно выполнить на ЧПУ, и практически любой вопрос, относящийся к станку или заданию. С течением времени ATHENA обучается, взаимодействуя с пользователями, постоянно улучшая свои навыки для помощи оператору. Такое взаимодействие в значительной степени повлияет на рост производительности.

Литература

1. Миронов С.Б. [Применение голосовых помощников и проблемы их использования в автоматизированном производстве]. – Научная статья, 2020 г.

2. Компания Макино [Электронный ресурс]. – Режим доступа: https://www.makino.eu/ru-ru/athena

Ю.В Ситникова, Д.С. Оплачко

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ АВТОРСТВА ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

(Самарский университет)

Введение

Рукописный текст использовался как средство документации и коммуникации на протяжении тысячи лет. Однако в наше время на замену рукописным текстам пришли электронные. За последнее десятилетие их количество резко возросло, и с каждым годом продолжает стремительно увеличиваться. Это связано, прежде всего, с широким распространением программ для обмена сообщениями в сети Интернет, возросшей роли электронной почты в деловой переписке, высокой популярности интернет-форумов, блогов и, конечно же, социальных сетей. Печатный текст, опубликованный в сети Интернет может содержать данные о том, кто его выложил, но не может дать исчерпывающей информации об авторе текста. Тут как раз и возникает вопрос о возможности определения авторства на основе текста и его содержания.

Определение авторства текста – это отнесение неидентифицированного текста на русском или других языках к одному из заранее перечисленных авторов. Определение авторства текста – это установление авторства неизвестных текстов, полагаясь на имеющуюся выборку авторов.

Определение автора анонимного текста в настоящее время является актуальной проблемой, так как идентификация авторства текста охватывает большой спектр целей: от отыскания автора необходимой статьи в интернете или запоминающегося отрывка художественного произведения до достаточно серьёзных военных и криминалистических целей.

Задача определения авторства текста является подвидом задачи классификации, так как относится только к массивам текстов и связана с разбиением их по авторам.

Классификация – один из разделов машинного обучения, посвященный решению следующей задачи. Имеется множество объектов (ситуаций), разделенных, некоторым образом, на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется обучающей выборкой. Классовая принадлежность остальных объектов не известна [1].

Формальная постановка задачи определения авторства текста

Рассмотрим формальную постановку задачи определения авторства. Она заключается в следующем: перед нами есть множество текстов (произведения или фрагменты произведений) $T = \{t_1, \dots, t_k\}$ на русском языке, а также множество авторов $A = \{a_1, \dots, a_l\}$. Для некоторого подмножество текстов $T' = \{t_1, \dots, t_m\} \subseteq T$ авторы известны, то есть существует множество пар «текст-автор» $D = \{(t_i, a_j)\}_{i=1}^{m}$, где для каждого элемента из подмножества установлен автор. Имеется неидентифицированных текст $t_i, i = \overline{m+1, k}, m \leq k$ из множества спорных или неидентифицированных (анонимных) текстов $T'' = \{t_{m+1}, \dots, t_k\} \subseteq T$ но известно, что он принадлежит кому-то из перечисленных выше авторов из множества А. Необходимо определить, кто из множества А является автором данного текста.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами. В этом случае множество *A* составляет множество предопределенных классов и их меток, *D* – обучающие примеры, а множество T'' – классифицируемые объекты. Целью является построение классификатора, решающего данную задачу, т. е. нахождение некоторой целевой функции $F: T \times A \to [-1,1]$, где −1 соответствует полностью отрицательному решению, 1 – положительному.

Нейронная сеть

Решить задачу классификации можно с помощью нескольких различных методов, одним из которых является метод с использованием нейронной сети.

Нейронная сеть (также искусственная нейронная сеть, ИНС) – математическая модель, а также её программное или аппаратное воплощение,

построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма [2].

В качестве архитектуры нейронной сети был выбран многослойный персептрон.

В настоящее время наиболее часто используемой архитектурой нейросети является многослойный персептрон (MLP), который представляет собой обобщение однослойного персептрона [3].

Многослойные персептроны успешно применяются для решения разнообразных сложных задач. При этом обучение с учителем выполняется с помощью такого популярного алгоритма, как алгоритм обратного распространения ошибки [3].

Проектирование автоматизированной системы определения авторства текста на естественном языке

В качестве архитектуры разрабатываемой системы была выбрана архитектура клиент-сервер.

В состав клиентской части входят следующие подсистемы:

1 Подсистема взаимодействия с сервером, которая осуществляет установку соединения с сервером, формирование и отправку запросов.

2 Подсистема визуализации, которая отображает пользовательский интерфейс.

1 Справочная подсистема, которая содержит сведения о системе (руководство пользователю) и об её разработчиках.

В состав серверной части входят следующие подсистемы:

1 Подсистема взаимодействия с клиентом, которая осуществляет приём данных с клиента и передачу их на обработку.

2 Подсистема подготовки текста, которая отвечает за процесс подготовки текста для классификатора. Она включает в себя:

- Подсистему предобработки текста, которая осуществляет токенизацию и нормализацию текста;
- Подсистему векторизации, которая преобразовывает текст в частотные векторы с помощью частотного анализатора и частотного словаря.

3 Подсистема классификации, которая отвечает за процесс классификации текстов. Она включает в себя:

- Подсистему обучения, которая отвечает за реализацию обучения нейронной сети;
- Подсистему настройки параметров классификатора, которая отвечает за ввод (выбор) значений параметров обучения и проверку корректности этих значений;
- Подсистему тестирования, которая отвечает за тестирование работы классификатора;
- Подсистему распознавания, которая отвечает за сопоставление загруженного текста с автором.

4 Файловая подсистема, которая отвечает за загрузку файла с текстом.

С помощью разрабатываемой системы пользователь сможет создать модель классификатора. Для этого ему потребуется подготовить обучающую выборку, настроить параметры классификатора и обучения, а также протестировать классификатор. Для подготовки обучающей выборки пользователь должен загрузить тексты для обучения, после чего обработать загруженный текст. По желанию пользователь может удалить из текста знаки пунктуации, стоп слова, а также использовать стеммер для нормализации текста. После этого будет проведен частотный анализ методом TF-IDF.

TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции [4].

Когда обучающая выборка будет подготовлена, пользователь может перейти к настройке параметров классификатора и обучения. Для этого ему будет необходимо ввести число нейронов в скрытом слое, коэффициент обучения, максимальное число итераций и допустимую ошибку. По заданным параметрам будет произведено обучение модели классификатора. Далее пользователь сможет протестировать созданную модель и сохранить её. Пользователь может также загрузить уже существующую модель из файла.

После создания модели классификатора пользователь сможет использовать её для определения авторства текста. Для этого пользователь должен выбрать одну из представленных моделей и загрузить текст для распознавания.

Заключение

Разработанная система была апробирована на обучающей выборке из 5000 текстов, относящимся к 10 классам (авторам). На основании этого были определены оптимальные значения параметров нейронной сети, при которых процент распознавания находится в диапазоне 69-75%. Для получения таких результатов количество нейронов на скрытом слое должно составлять 100, а коэффициент обучения должен быть равен 0,01.

Литература

1. Классификация [Электронный ресурс] // machinelearning.ru: [сайт]. URL: http://www.machinelearning.ru/wiki/index.php?title=Классификация (дата обращения: 20.03.2022).

2. Искусственная нейронная сеть [Электронный ресурс] // dic.academic.ru: [сайт]. URL: https://dic.academic.ru/dic.nsf/ruwiki/13889 (дата обращения: 23.03.2022).

3. Солдатова, О.П. Основы нейроинфоматики [Текст] : учеб. пособие / О.П. Солдатова. – Самара: Изд-во Самар, гос. аэрокосм, ун-та, 2006. – 132 с. : ил.– ISBN 5-7883-0467-9.

4. TF-IDF [Электронный ресурс] // ru.wikipedia.org: [сайт]. URL: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения: 23.03.2022).