УДК 004

РАСПОЗНАВАНИЯ РУКОПИСНЫХ ТЕКСТОВ ДЛЯ ЭКСПЕРТНОЙ ОЦЕНКИ ЗНАНИЙ ПО РЕЗУЛЬТАТАМ ИТОГОВОЙ АТТЕСТАЦИИ

Я.И. Казюлина
Научный руководитель: А.В. Лепустин, старший преподаватель каф. ВТ ИК ТПУ
Национальный исследовательский Томский политехнический университет
E-mail: kyai@tpu.ru

This article is about the creation of the special software which will recognize the handwriting for the purpose of the expert knowledge assessment of the final certification results.

Keywords: Optical character recognition, or OCR, Method of recognition, neural Networks, library Puma.net.

Автору статьи необходимо создать специальный модуль, которая будет распознавать бланки ЕГЭ, сверять полученные знания с правильными ответами и сохранять все данные в базу данных.

Оптическое распознавание символов (англ. optical character recognition, OCR) – механический или электронный перевод изображений рукописного, машинописного или печатного текста в текстовые данные – последовательность кодов, использующихся для представления символов в компьютере (например, в текстовом редакторе). Распознавание широко используется для конвертации книг и документов в электронный вид, для автоматизации систем учёта в бизнесе или для публикации текста на веб-странице.

Написание приложения будет происходить на языке C#, т. к. C# – один из наиболее широко используемых языков программирования в мире. Хорошо написанные программы на C# работают быстро и эффективно. C# является более гибким, чем другие языки, поддерживает функциональное и объектно-ориентированное программирование.

Ввиду, того что часть программы уже написано, средой разработки клиентского приложения выбрана – Microsoft Visual Studio.

На сегодняшний момент библиотек распознавания текста создано не мало, но большинство из них платные, а автору в работе понадобятся бесплатные библиотеки, поэтому список библиотек существенно сократился.

Перечень библиотек изученные автором: Tesseract, Puma.net, AForge, GOCR, OpenCV. Выбор был сделан в сторону библиотеки Puma.net.

Puma.NET представляет собой оболочку для библиотеки распознавания Cognitive Technologies CuneiFrom, которая позволяет легко включать функций распознавания в любом NET Framework 2.0 (или выше) приложении. API предоставляется через ряд простых классов. Высокая производительность и точность результатов распознавания может быть достигнута с помощью пары строчек кода.

На рис. 1 и 2 представлен результат распознавания, используя библиотеку Puma.NET.

На данном примере рассмотрим пример распознавания только цифр. Рассмотрим еще пример работы программы, где будут использоваться русские буквы.

Бланк ЕГЭ с русскими буквами и результат распознавания представлен на рис. 3 и 4.

Анализируя полученные результаты, можно сделать вывод, что результат распознавания цифр – 60 %. Результат распознавания букв 100 %. Следует улучшать распознавание цифр.

Для улучшения распознавания, автором был выбран путь-создание нейронной сети, распознающей символы.

Нейронная сеть или нервная система человека – это сложная сеть структур человека, обеспечивающая взаимосвязанное поведение всех систем организма.

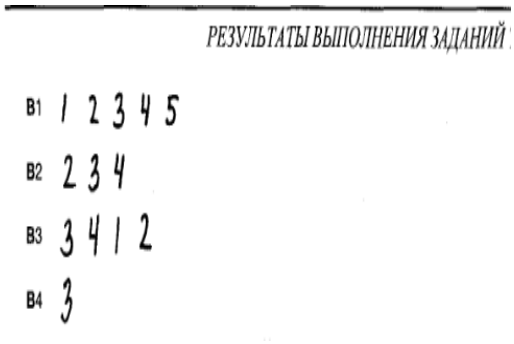

Рис. 1. Сканированный бланк ЕГЭ

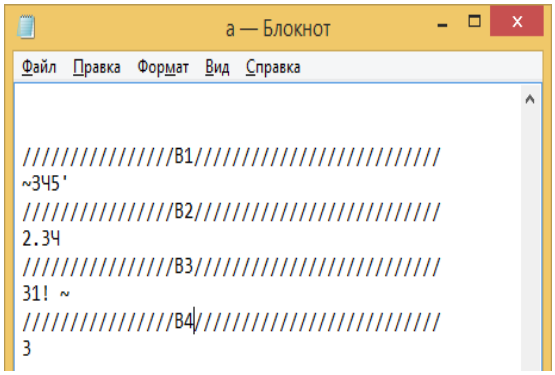

Рис. 2.Распознаные данные

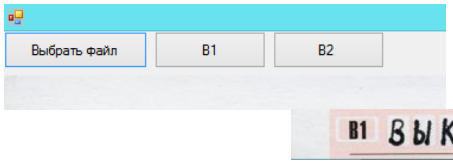

Рис. 3. Вырезание части B1 из общего бланка

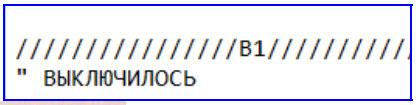

Рис. 4. Результат работы программы

Нейронные сети в искусственном интеллекте – это упрощенные модели биологических нейронных сетей.

У нейронных сетей много важных свойств, но ключевое из них – это способность к обучению. Обучение нейронной сети в первую очередь заключается в изменении «силы» синаптических связей между нейронами.

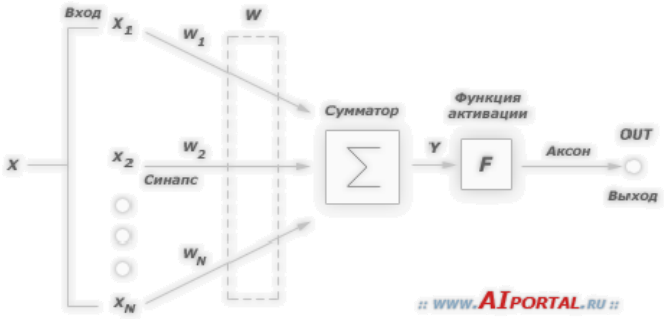

Рис. 5. Модель нейрона

Функция активации выбрана-логсигмоидная.

Автором было создано приложение позволяющее создавать, сохранять нейронную сеть, создавать обучающую тестирующую выборку, открывать ранее созданную нейронную сети. Созданное приложение на основе нейронных сетей 100 % распознает рукописные цифры.

Для работы автору не обходимо хранить бланки ЕГЭ и полученные результаты в БД. Будем использовать средства управления базой данных(СУБД) – Microsoft SQL Server Management Studio 2008 R2.

Изображение в базе данных будем хранить так: при запросе к ней мы в одном из полей выборки получаем байтовый массив, который и является самим изображением. Важно отметить, что этот способ хранения лучше всего использовать для небольших картинок.

Список литературы

1. Ян Д.Е., Анисимович К.В., Шамис А.Л. Новая технология распознавания символов. Теория, практическая реализация, перспективы. – М.: Препринт, 1995.

2. Промахина И.М., Коростелев А.П. Об одном классе вероятностных рекуррентных алгоритмов распознавания. – М.: Препринт, 1984.
3. Y-H Pao Adaptive pattern recognition and neural network “Addison-Wesley” 1989.
4. Puma.net: [Электронный ресурс]. Project Description M., 1997–2014. URL: http://pumanet.codeplex.com/ (Дата обращения: 18.10.2014).

УДК 004

ПРОГНОЗИРОВАНИЕ СОСТОЯНИЯ ФОНДОВОГО РЫНКА НА ОСНОВЕ ФИНАНСОВЫХ НОВОСТЕЙ

Г.Г. Петрова
Научный руководитель: А.Ф. Тузовский, д.т.н., профессор каф. ОСУ ИК ТПУ
Национальный исследовательский Томский политехнический университет,
634050, Томск, пр. Ленина, 30
E-mail: ggp_pgg@mail.ru

Abstracts. *In this paper are discussed existing methods of Semantic Web technologies application for financial news processing.*

Keywords: Semantic Web, ontology, financial news, stock market.
Ключевые слова: Семантическая Паутина, онтология, финансовые новости, фондовый рынок.

Введение

В современном информационном мире всё больше специалистов фондового рынка пользуются новостными информационными порталами в сети Интернет. Содержание этих порталов отражает прошлые, текущие и будущие события в мире, что является ценной информацией для различного прогнозирования. Инвесторы и трейдеры фондового рынка используют знания о текущей мировой ситуации для принятия решений о покупке или продаже ценных бумаг. Одним из самых важных источников информации являются финансовые новости. Объем и скорость изменения финансовых новостей от различных источников увеличиваются, что делает сложным и трудоемким их обработку вручную в условиях фондового рынка.

Такие новости используются (учитываются) при формировании решений экспертов финансовой области. Например, при торговле ценными бумагами, формировании инвестиционного портфеля и т. п. В связи с этим имеется большая потребность в автоматизации извлечения и обработки информации финансовых новостей из сети Интернет [1].

Извлечение информации из финансовых новостей

Для извлечения и хранения финансовых новостей в виде, пригодном для машинной обработки большую помощь могут оказать технологии Semantic Web [2]. Для классификации финансовых новостей могут использоваться онтологии – формализованное представление некоторой области знаний, включающее в себя определение понятий, их свойств и взаимосвязей между ними. Для описания онтологий используют язык OWL (Web Ontology Language). Заполнение онтологий происходит на основе шаблонов: лексико-семантических и лексико-синтаксических [3]. *Лексико-синтаксические шаблоны* – это характерные выражения (словосочетания и обороты), конструкции из определенных элементов языка. *Лексико-семантические шаблоны* представляют собой фреймовую структуру, где слотами фреймов являются семантические роли, а значения слотов – слова естественного языка, взятые из синтаксического анализа предложения. Данные шаблоны позволяют выделять большее количе-