УДК 004.912
DOI: 10.25559/SITITO.14.201804.872-880

ОБЛАЧНЫЕ СЕРВИСЫ ДЛЯ ОБРАБОТКИ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Р.И. Мухамедиев[1,2,3], А. Сымагулов[1,2], Я.И. Кучин[1], С. Абдуллаева[1], Ф.Н. Абдолдина[2]
[1] Институт информационных и вычислительных технологий МОН РК, г. Алматы, Казахстан
[2] Казахский национальный исследовательский технический университет имени К.И. Сатпаева, г. Алматы, Казахстан
[3] Высшая школа менеджмента и информационных систем, г. Рига, Латвия

CLOUD SERVICES FOR NATURAL LANGUAGE PROCESSING

Ravil I. Mukhamediev[1,2,3], Adilkhan Symagulov[1,2], Yan I. Kuchin[1], Sabina Abdullayeva[1], Farida N. Abdoldina[2]
[1] Institute of Information and Computational Technologies MES RK, Almaty, Kazakhstan
[2] Kazakh National Research Technical University named after K.I. Satpaev, Almaty, Kazakhstan
[3] ISMA University, Riga, Latvia

Ключевые слова

Обработка естественного языка (ОЕЯ); обработка естественно-языковых текстов; облачные сервисы; программное обеспечение как сервис; морфологическая разметка; распознавание именованных сущностей; тональность; категоризация текстов; Accuracy; Precision; Recall; гармоническое среднее.

Аннотация

В работе приведены результаты экспериментов, проведенных с целью сравнительного анализа качества работы существующих облачных сервисов по обработке текстов на русском языке. В статье приводится обзор 10-ти облачных сервисов: TextRazor, RosetteTextAnalytics, EurekaEngine, CloudNaturalLanguage, Texterra, Pullenti, NER-ru, UDPipe, AOT, DeepPavlov. Количественные исследования качества работы выполнены для 6-ти из них. В процессе оценки сервисов анализировалось выполнение таких функций, как частеречная разметка, анализ тональности, распознавание именованных сущностей и категоризация текстов. Для сравнительной оценки качества работы сервисов были использованы материалы соревнований: factRuEval-2016 (выявление именованных сущностей), AlemResearch (тональность) и корпуса текстов, Taiga и OpenCorpora (морфологическая разметка). Качество распознавания именованных сущностей оценивалось путем расчета параметров Accuracy, Precision, Recall и F1.

В результате проведенного исследования показано, что при решении задач обработки текста на русском языке для распознавания именованных сущностей и определения тональности текста наилучший результат показывает сервис EurekaEngine, а для частеречной разметки текста – сервис RosetteTextAnalytics, для категоризации текста – сервис TextRazor.

Об авторах:

Мухамедиев Равиль Ильгизович, доктор инженерных наук, профессор, главный научный сотрудник, Институт информационных и вычислительных технологий МОН РК (050010, Республика Казахстан, г. Алматы, ул. Пушкина, д. 125); профессор, кафедра программная инженерия, Казахский национальный исследовательский технический университет имени К.И. Сатпаева (050013, Республика Казахстан, г. Алматы, ул. Сатпаева, д. 22а); профессор, Dr.Sc., Высшая школа менеджмента информационных систем (LV-1019, Латвия, г. Рига, ул. Ломоносова, д. 1, корп. 6), ORCID: http://orcid.org/0000-0002-3727-043X, ravil.muhamedyev@gmail.com

Сымагулов Адилхан, бакалавр естественных наук, инженер-программист, Институт информационных и вычислительных технологий МОН РК (050010, Республика Казахстан, г. Алматы, ул. Пушкина, д. 125); магистрант, Казахский национальный исследовательский технический университет имени К.И. Сатпаева (050013, Республика Казахстан, г.Алматы, ул. Сатпаева, д. 22а), ORCID: http://orcid.org/0000-0001-9974-3215, asmogulove00@gmail.com

Кучин Ян Игоревич, магистр, научный сотрудник, Институт информационных и вычислительных технологий МОН РК (050010, Республика Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0002-5271-9071, ykuchin@mail.ru

Абдуллаева Сабина, бакалавр математических методов в экономике, лаборант, Институт информационных и вычислительных технологий МОН РК (050010, Республика Казахстан, г. Алматы, ул. Пушкина, д. 125), ORCID: http://orcid.org/0000-0002-2212-0500, Sabina.abdullaeva23@gmail.com

Абдолдина Фарида Наурузбаевна, кандидат технических наук, заместитель директора, Институт информационных и телекоммуникационных технологий, Казахский национальный исследовательский технический университет имени К.И. Сатпаева (050013, Республика Казахстан, г. Алматы, ул. Сатпаева, д. 22а), ORCID: http://orcid.org/0000-0003-1816-6343, farida_mail@mail.ru

Keywords

Natural language processing (NLP); Word processing; Cloud services; Software as a service (SaaS); Natural language texts processing; Part of speech tagging; Named entity recognition; Sentiment analysis; Text categorization; Accuracy; Precision; Recall; Harmonic average.

Abstract

The paper presents the results of experiments conducted with the aim of a comparative analysis of the performance of the existing cloud services for natural language processing in Russian. The article provides an overview of 10 cloud services: TextRazor, RosetteTextAnalytics, EurekaEngine, CloudNaturalLanguage, Texterra, Pullenti, NER-ru, UDPipe, AOT, DeepPavlov. Quantitative studies of their performance were made for 6 of them. In the process of evaluating services, the execution of such functions as the part of speech tagging, sentiment analysis, named entity recognition and the categorization of texts were analyzed. For a comparative assessment of the quality of the services, the following competition materials were used: factRuEval-2016 (named entities), AlemResearch (sentiment) and the corpora, Taiga and OpenCorpora (part of speech). The named entities recognition quality was evaluated by calculating Accuracy, Precision, Recall, and F1 parameters.

As a result of the study, it was shown that when solving natural language text processing tasks in Russian, the best result is shown by the EurekaEngine service for recognizing named entities and sentiment analysis of the text, RosetteTextAnalytics service proved best in part of speech tagging the and TextRazor service in text categorization.

Введение

Обработка естественного языка (Natural Language Processing – NLP) бурно развивающаяся область исследований, результатами которых в виде технологий обработки речи и текстов мы активно пользуемся. Задача этого направления – разработка методов автоматического анализа и представления человеческого языка [1]. Данное направление решает широкую гамму прикладных задач, к числу которых относятся: автоматический перевод [2]; автоматическое реферирование; генерация ответов на запросы пользователя (Question Answering) [3] в том числе, Visual Question Answering, когда по запросу пользователя на естественном языке требуется найти элементы изображения или извлечь общую информацию об изображении [4]; извлечение информации (Information Extraction)– преобразование неструктурированной информации естественно-языковых текстов в структурированную форму [5], в частности, «свежее» направление Open IE, который не ограничивается небольшим набором целевых отношений типа {аргумент1; отношение; аргумент2} известных заранее, а предполагает извлечение всех типов отношений из разнородных корпусов с использованием кластерных (Unsupervised) методов или полу-управляемого обучения (Semi-supervised) [6]; информационный поиск (Information Retrieval) – процесс получения и представления наиболее соответствующей запросу информации из большой коллекции информационных ресурсов [7, 8], в том числе временной [9], географической [10], событийной и др.; анализ тональности [11] и другие направления так или иначе связанные с обработкой устной и письменной естественной речи.

Потребности в решении практических задач NLP послужили катализатором развития методов, среди которых можно упомянуть: Machine Learning, Neural networks, Deep Learning [12], Fuzzy logic, Firstorder logic, Semantic networks и другие.

В последнее время научные исследования и общий возросший уровень вычислений привел к ряду прорывных результатов в NLP, среди которых можно отметить успехи в области автоматического перевода (machine translation), реферирования (automatic summarization), информационного поиска (information retrieval), вопросно-ответных систем (question answering), анализа тональности (sentiment analysis), извлечения информации (information extraction [13]), авторской верификации (authorship verification [14]).

Ключевыми аспектами, позволившими получить впечатляющие результаты в области автоматической обработки естественно-языковых текстов, являются, по мнению [15], успехи в развитии методов машинного обучения, многократное возрастание вычислительных мощностей, доступность большого объема лингвистических данных и развитие в понимании структуры естественного языка в приложении к социальному контексту. Со времен работы [16] достигнуты впечатляющие результаты в области разработки статистических моделей языка. Большие объемы доступных текстов социальных медиа и применение глубоких нейронных сетей [17] приводят к постановке задач извлечения закономерностей из огромных объемов неструктурированной информации на основе успехов в дистрибутивной лингвистике и так называемого distantly supervised learning [18].

Можно отметить, что задачи «инфраструктурного» типа, относящиеся к области морфологического и синтаксического анализа, статистической модели языка, в значительной мере решены и могут быть кооптированы в разрабатываемую систему, например в форме программного обеспечения как сервиса (software as a service –SaaS) [19, 20].

Разработчики, реализующие парадигму SaaS, предоставляют полностью настроенное программное обеспечение (ПО), развернутое на собственных серверах, что позволяет им иметь полный контроль над приложением. Основным достоинством SaaS для клиентов является отсутствие необходимости поддерживать работу ПО, выделять свои вычислительные мощности, хранилища, оборудование. Можно сказать, что клиенты арендуют ПО и все сопутствующие технологии для решения своих задач. Наличие месячных подписок и определенных ценовых политик позволяет пользоваться ПО тогда, когда нужно и оплачивать только то, что использовалось. Создателям, в свою очередь, не нужно опасаться пиратского использования их продукта, потому что исходные коды продукта хранятся только у его разработчиков и не отправляются клиентам.

В настоящее время облачные сервисы для обработки текстов на естественном языке представляют собой весьма вос-

требованное и интенсивно развивающееся направление услуг. Вместе с тем, для выбора сервиса недостаточно рекламной информации. Как говорится, “Одно точное измерение стоит мнений тысяч экспертов” (“One accurate measurement is worth more than a thousand expert opinions”)[1]. Таким образом, решено было провести самостоятельные эксперименты по анализу работы сервисов обработки текста, результаты которых изложены в настоящей статье.

Работа состоит из следующих частей:

- в первом разделе описываются анализируемые облачные сервисы NLP;
- во втором разделе мы описываем использованные текстовые корпуса, метрики оценки и приводим пример их расчета;
- в третьем разделе описываются полученные результаты сравнительного анализа сервисов;
- в заключении подводятся итоги сравнения и намечаются задачи для дальнейших исследований.

Облачные сервисы для NLP

В настоящей работе нас интересуют задачи обработки естественно-языковых текстов (ОЕЯТ) и их решение с помощью SaaS. Анализ имеющихся сервисов ОЕЯТ необходим для уточнения границ их функциональных возможностей и качества выполнения имеющихся функций. Немаловажным является также понимание политики использования этих сервисов. Существование облачных сервисов для ОЕЯТ позволяет исследователям сэкономить время и силы на создании, развертывании и поддержке необходимых технологий для исследований в области NLP. Облачные сервисы ОЕЯТ предлагают такие возможности, как распознавание именованных сущностей, частеречная разметка, анализ тональности, стемминг, лемматизация, категоризация, которые, как отмечено выше, являются основой для решения прикладных задач NLP [21, 22, 23].

В данной работе рассмотрены возможности следующих систем: TextRazor, RosetteTextAnalytics, EurekaEngine, CloudNaturalLanguage, Texterra, Pullenti, NER-ru, UDPipe, AOT, DeepPavlov.

Для проведения анализа использованы небольшие размеченные корпуса текстов[2] и выполнен подсчет точностных показателей работы сервисов.

Вначале кратко опишем функциональные возможности, политику использования и особенности анализируемых сервисов ОЕЯ.

TextRazor сочетает технологии NLP с базой знаний о реальных фактах (DBpedia и Freebase). Для русского языка способен выделять категории текста, производить токенизацию и стемминг, определять именованные сущности. Не способен производить частеречную разметку текста и лемматизацию. Метод распознавания именованных сущностей - базы данных DBpedia и Freebase. Для английского языка есть возможность производить частеречную разметку, лемматизацию, определение именованных сущностей, связей слов и предположения о значении слов в контексте. Отсутствует тональная разметка текста. Политика использования: бесплатный план использования предполагает весь функционал и 500 запросов в день, большее количество запросов в день можно получить путем приобретения одной из месячных подписок. Даются скидки для исследователей. Уникальность TextRazor заключается в том, что определенные именованные сущности имеют ссылку на соответствующие страницы в базах данных Freebase и Wikipedia, а так же в подробном выделении категорий текста. Сервис имеет подробное руководство по использованию API (программный интерфейс приложения)[3].

RosetteTextAnalytics использует технологии NLP для поиска информации, интеллектуального анализа текста, машинного обучения, статистики и вычислительной лингвистики. Предоставляет такие возможности, как определение имен, мест и других ключевых понятий из тела неструктурированного текста. RosetteTextAnalytics способна выполнять частеречную разметку текста, лемматизацию, токенизацию, определение именованных сущностей. Метод распознавания именованных сущностей - статистические алгоритмы, нейронная сеть и списки именованных сущностей. Для английского языка имеет возможность производить категоризацию и тональную разметку текста. Отсутствует стемминг. Политика использования**:** облачные планы использования включают в себя бесплатный план, который позволяет производить до 10 000 API-запросов в месяц, 1000 запросов в день. Планы от 40 000 запросов без ограничения месяц/день имеют цены от 100$ в месяц и выше при увеличении количества API-запросов. Уникальность сервиса заключается в свободном доступе к частеречной разметке и лемматизации текста для русского языка. Сервис имеет примеры использования API для разных языков. Предоставляет API-ключ для идентификации пользователя, а также настроенные API привязки для разных языков программирования [4].

EurekaEngine — российская система, которая способна извлекать значения и факты из текстов. Обеспечивается категоризация текста, тональная разметка, выявление именованных сущностей. Метод распознавания именованных сущностей - статистические алгоритмы. Ограниченная демо-версией частеречная разметка, лемматизация и стемминг - обрабатывает только первые 10 слов текста. Политика использования: возможно использование демо-версии, при необходимости использования платной версии нужно будет потратить 800 евро за 50 000 документов в месяц, 2400 евро за 300 000 документов в месяц и далее по возрастанию количества документов будет расти и стоимость. Уникальность заключается в наличии тональной разметки текста, по сравнению с TextRazor и RosetteTextAnalytics. Доступ к API только после оплаты[5].

CloudNaturalLanguage от Google — предоставляет API для построения веб сервисов, представляющий собой клиент-серверное решение, построен на http запросах. Поддерживает возможности обработки и анализа текста с использованием готовых моделей. Возможно извлечение именованных сущностей

1 Фразу связывают с именем контр-адмирала ВМФ, ученого, создательницы первого в мире компилятора языка программирования Грейс Хоппер.
2 Небольшие размеченные корпуса текстов. [Электронный ресурс]. URL: https://drive.google.com/open?id=1YWaFOiGBXT_c9ueH3J0BhI2G3PcZXb39 (дата обращения: 12.09.2018).
3 TextRazor. [Электронный ресурс]. URL: https://www.textrazor.com/about (дата обращения: 12.09.2018).
4 RosetteTextAnalytic. [Электронный ресурс]. URL: https://www.rosette.com/ (дата обращения: 12.09.2018).
5 EurekaEngine. [Электронный ресурс]. URL: http://eurekaengine.ru/en/description/ (дата обращения: 12.09.2018).

из текста, частеречная разметка со связями и зависимостями, тональная разметка, категоризация текста. Для использования данного сервиса, как и других google сервисов, нужно создать проект на сайте. Затем, получив API ключ, использовать его в клиентских HTTP запросах. Цена за текст зависит от количества символов в нем и варьируется от 2 долларов до 25 центов. Анализ текста до 5000 символов бесплатный. Сервис поддерживает 9 языков (китайский, английский, французский, немецкий, итальянский, японский, корейский, португальский, испанский). Тексты на языках, которые не поддерживаются, можно перевести с помощью другого облачного сервиса от google — Translate API. Cloud NLP имеет подробную документацию и примеры использования на различных языках программирования [6].

Texterra — создан командой отдела информационных систем, института системного программирования им. В.П. Иванникова РАН. Для решения задач NLP использует статистические и лингвистические методы. Сервис обеспечивает морфологический и синтаксический анализ, основанный на синтаксически размеченном корпусе русского языка (Синтагрус), распознавание именованных сущностей, возможность поиска слов из текстов в Википедии. Метод распознавания именованных сущностей - машинное обучение. Политика использования: можно пользоваться демо-версией продукта, а для того, чтобы получить API-ключи, нужно зарегистрироваться. Уникальность заключается в том, что сервис выполняет синтаксический анализ основанный на корпусе Синтагрус. Для разработчиков предоставляются SDK (software development kit) на двух языках программирования: Python и Ruby, а для всех остальных онлайн-демо [7].

Pullenti — предоставляет распознавание именованных сущностей, построение семантической сети на основе синтаксического анализа всего текста, развернутый морфологический анализ (часть речи, род, падеж, число), а также множество доступных вариантов морфологического разбора слов в тексте. Распознавание именованных сущностей основано на правилах и словарях сущностей. Политика использования: бесплатный для некоммерческого использования. Уникальность сервиса заключается в наличии семантического анализа текста. Имеется SDK для .NET, а так же для таких языков программирования как Java и Python [8].

NER-ru — сервис, созданный человеком или группой под псевдонимом Zamgi на GitHub (сервисе хостинга IT проектов, основанном на контроле версий). Позволяет производить распознавание именованных сущностей. Метод распознавания именованных сущностей - статистические алгоритмы. Политика использования: бесплатный для некоммерческого использования. Уникален своим минимализмом. Весь проект находится на GitHub, API отсутствует [9].

UDPipe — это чешское ПО, разработанное в институте формальной и прикладной лингвистики, Карлов университета. Позволяет производить развернутый морфологический анализ (часть речи, род, падеж, число), токенизацию, лемматизацию, а также строить деревья зависимостей между словами. Методы, используемые для решения задач обработки естественно-языковых текстов - банки деревьев из открытого сообщества по кросс-лингвистической разметке текста UniversalDependencies. Политика использования: бесплатный для некоммерческого использования. Уникальность сервиса заключается в возможности строить деревья зависимостей для каждого предложения в тексте [10].

AOT — российское ПО для решения задач обработки естественно-языковых текстов. Позволяет производить синтаксический анализ предложений, основанный на правилах русского языка, а так же частеречную разметку, но только в веб версии, так как проект больше не поддерживается. Политика использования: бесплатный для некоммерческого использования. Уникальность заключается в синтаксическом анализе предложений, основанном на правилах русского языка [11].

DeepPavlov — библиотека диалоговых систем, разрабатываемая в лаборатории Московского физико-технического института. Библиотека содержит модули для обработки русского и английского языков. В основе методов обработки текста лежат глубокие нейронные сети. Данная библиотека также реализована в виде сервиса iPavlov, где также присутствует возможность анализа текстов. Весь исходный код библиотеки размещен в открытом доступе [12].

Анализ распознавания именованных сущностей с помощью сервиса DeepPavlov был проведен на данных, полученных Институтом информационных и вычислительных технологий МОН РК с сайта Tengrinews. Другие сервисы работали с корпусом текстов factRuEval-2016, поэтому результаты DeepPavlov не показаны в таблицах сравнения.

В таблице 1 приведены характеристики исследуемых облачных сервисов.

Таблица 1. Характеристики облачных сервисов для обработки текстов на естественном языке
Table 1. Characteristics of cloud services for natural language processing

Функции \ Наименование сервиса	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru	UDPipe	AOT	CloudNaturalLanguage
Частеречная разметка	Для рус. яз отсутствует, для англ. яз есть	Есть	-	Есть Рус. яз - Синтагрус Англ. яз. - Penntreebank	Есть Имеет: варианты слова, число, род, падеж	-	Есть Основана на UDTreebanks	Есть Основана на словаре А.А. Зализняка	Есть
Токенизация	Есть	Есть	Есть	Есть	Есть	-	Есть	Есть	Есть

6 CloudNaturalLanguage. [Электронный ресурс]. URL: https://cloud.google.com/natural-language/ (дата обращения: 12.09.2018).
7 Texterra. [Электронный ресурс]. URL: https://api.ispras.ru/demo/texterra (дата обращения: 12.09.2018).
8 Pullenti. [Электронный ресурс]. URL: http://pullenti.ru/DemoPage.aspx (дата обращения: 12.09.2018).
9 NER-ru. [Электронный ресурс]. URL: http://ner-ru.apphb.com/index.html (дата обращения: 12.09.2018).
10 UDPipe. [Электронный ресурс]. URL: http://lindat.mff.cuni.cz/services/udpipe/ (дата обращения: 12.09.2018).
11 AOT. [Электронный ресурс]. URL: http://aot.ru/history.html (дата обращения: 12.09.2018).
12 DeepPavlov. [Электронный ресурс]. URL: https://deeppavlov.ai/ (дата обращения: 12.09.2018).

Функции \ Наименование сервиса	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru	UDPipe	AOT	CloudNaturalLanguage
Стемминг	Есть	-	-	-	-	-	-	-	-
Лемматизация	Для рус. яз отсутствует, для англ. яз есть	Есть	-	Есть	Есть	-	Есть	Есть	Есть
Анализ тональности	-	-	Есть	Есть	-	-	-	-	Есть
Категори-зация	Есть	Есть для англ. языка	Есть	-	-	-	-	-	Есть
Ценовая политика	Демо, 6000 API запросов - 200$ в месяц	Демо, 40 000 API запросов - 100$ в месяц	Демо, 50 000 документов - 800 евро в месяц	Бесплатно	Бесплатно для некоммерческого использования	-	Бесплатно для некоммерческого использования	Бесплатно	Зависит от количества символов в тексте
Наличие демо	Есть	Есть	Есть	Есть	Есть	Есть	Есть	Есть	Есть
Количество поддержи-ваемых языков	11	20 для распозн. именов. сущн., 32 для частеречной разметки	1 (рус)	2 (рус, англ)	3 (рус, укр, англ)	1 (рус)	74	3 (рус, англ, нем)	9
Уни-каль-ность	Сущности со ссылками на Freebase и Wikipedia	Стабильная частеречная разметка и лемматизация	Наличие тональной разметки текста	Морфологический анализ основанный на корпусе СинТагРус	Семантический анализ текста	Минимализм	Построение Синтаксических деревьев	Синтаксический анализ предложений	-

Метод исследования

В процессе оценки сервисов анализировалось качество выполнения таких функций, как *частеречная разметка, анализ тональности, распознавание именованных сущностей, категоризация текстов.*

Для сравнения сервисов между собой были использованы следующие три небольших корпуса текстов. Корпусом для определения качества выявления именованных сущностей послужили материалы с ежегодного соревнования Dialogue Evaluation factRuEval 2016, на котором проводятся сравнительные тестирования систем компьютерного анализа, решающих, в том числе, задачу выявления именованных сущностей [13]. Сущности в текстах выделялись вручную. Для определения качества работы сервисов использовано десять размеченных текстов (среднее количество слов в каждом тексте 143). На вход сервисов подавались указанные тексты, затем, определенные сервисами сущности сравнивались с ручной разметкой.

Указанный микро корпус использовался и для оценки качества категоризации.

Для определения качества частеречной разметки использовались тексты из коллекции FakeNews, принадлежащей корпусу Taiga [24]. Использованы три текста (157, 188, 180 слов), морфологическая разметка которых выполнена с использованием пакета UDPipe. На вход сервиса подавался текст со снятой разметкой, затем проводилось сравнение с исходной разметкой.

Корпусом для определения качества тональной разметки текста послужили тексты с соревнования Artificial Intelligence Day, организованного компанией Alem Media Monitoring (AlemResearch) и Казахско-Британским техническим университетом. Для определения качества тональной разметки текста использовано 20 текстов (10 негативных со средним количеством слов в одном тексте 238 и 10 позитивных со средним количеством слов в одном тексте—313). На вход сервиса подавались указанные тексты, затем рассчитывалось количество ошибок сервиса.

Для количественной оценки распознавания именованных сущностей использованы параметры Accuracy (A), Precision (P), Recall (R) и гармоническое среднее F1, которые повсеместно применяются для оценки алгоритмов и программ классификации данных [25]. Расчет указанных параметров базируется на следующих количественных показателях:

- количество правильно распознанных сущностей (tp — истинно положительные);
- количество сущностей, которые неверно выделены как сущности (fp — ложно-положительные);
- количество реально имеющихся сущностей, не распознанных сервисом (fn — ложно-отрицательные).

Отметим, что в задачах оценки алгоритмов классификации используется показатель tn равный количеству объектов противоположного класса правильно распознанных классификатором. Поскольку количество таких объектов неизвестно, то данный показатель во всех случаях принят равным 0. Исходя из принятых обозначений:

$$Accuracy = (tp+tn)/(tp+tn+fp+fn) \quad (1)$$

показывает долю правильно распознанных сущностей к общему количеству сущностей. Ошибка первого рода:

$$Precision = tp/(tp+fp) \quad (2)$$

показывает точность сервиса, то есть насколько часто он приписывает лишние сущности к имеющимся. Чем выше Precision сервиса, тем меньше лишних сущностей он определяет. Ошибка второго рода:

$$Recall = tp/(tp+fn) \quad (3)$$

показывает, насколько полно сервис определяет имеющиеся сущности. Чем выше Recall сервиса, тем меньше пропусков в распознавании именованных сущностей он допускает.

F-мера (F1), представляет собой гармоническое среднее между точностью (Precision) и полнотой (Recall). Часто вычисляется с помощью следующего выражения

13 *Starostin A S.* factRuEval 2016: evaluation of named entity recognition and fact extraction systems for Russian // Dialogue 2016. [Электронный ресурс]. URL: http://www.dialog-21.ru/dialogue2016/results/ (дата обращения: 12.09.2018).

F_1=2* (precision*recall)/(precision+recall) (4)

Рассмотрим пример расчета показателей качества выявления именованных сущностей

Один из текстов, содержащийся в корпусе текстов гласит: «Встреча с послом Италии в миде Грузии. По инициативе итальянской стороны чрезвычайный и полномочный посол Италии в Грузии Витторио Сандали встретился с заместителем министра иностранных дел Грузии Александром Налбандовым. Предметом обсуждения стали вопросы сотрудничества в международных организациях».

В данном тексте из корпуса factRuEval 5 отмеченных вручную именованных сущностей («Италия», «МИД Грузии», «Грузия», «Витторио Сандали», «Александр Налбандов»).

TextRazor правильно определил 3 сущности. Вместе с тем, всего определил 4 сущности («Италия», «Грузия», «Посол», «Витторио Сандали»), пропустив две из размеченных вручную («МИД Грузии», «Александр Налбандов»). Таким образом, tp = 3, fp = 1, fn = 2, что в соответствии с формулами 1-4 дает следующий результат: A = 3/(3+0+1+2) = 0,5; P = 3/(3+1) = 0,75; R = 3/(3+2) = 0,6; F1 = (2 * ((0,75 * 0,6) / (0,75 + 0,6))) = 0,66.

Анализ качества категоризации выполнялся относительно субъективно. Авторами статьи были рассмотрены 10 упомянутых текстов, вручную выделены категории текстов.

Результаты анализа сервисов

Показатели качества по распознаванию именованных сущностей

Результаты анализа сервисов по распознаванию именованных сущностей представлены в таблицах 2-6. Усредненные результаты по каждому показателю для всех сервисов представлены в таблице 7, для наглядности результаты так же представлены в виде графика на рисунке 1. Согласно проведенным экспериментам, наилучший результат по исследуемым показателям показал сервис EurekaEngine (таблица 7).

DeepPavlov на упомянутом выше корпусе показал: Precision = 0,88, Recall = 0,75, F1 = 0,82.

Таблица 2. Результаты работы сервисов по показателю Accuracy

Table 2. The results in terms of Accuracy

Текст №	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru
1	0,5	0,44	0,8	0,66	1	0,8
2	0,57	0,83	0,83	1	0,42	0,83
3	0,28	0,75	1	0,4	1	0,5
4	0,78	0,92	0,92	0,35	0,62	0,6
5	0,58	0,78	1	0,92	0,69	0,92
6	0,5	0,57	0,89	0,73	0,46	0,8
7	0,33	0,41	0,71	0,42	0,56	0,41
8	0,22	0,75	1	0,92	0,11	0,64
9	0,37	0,53	0,61	0,47	0,53	0,54
10	0,70	0,62	0,86	0,67	0,93	0,71

Таблица 3. Результаты работы сервисов по показателю Precision

Table 3. The results in terms of Precision

Текст №	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru
1	0,75	0,5	1	0,8	1	1
2	0,67	0,83	0,83	1	0,6	0,83
3	0,33	0,75	1	0,5	1	0,5
4	0,91	1	1	0,83	0,77	0,81
5	0,62	0,78	1	0,92	0,81	0,92
6	0,5	0,61	1	0,8	0,6	0,89
7	0,58	0,7	1	1	0,7	0,7
8	0,27	0,75	1	1	0,17	0,82
9	0,45	0,62	1	0,78	0,69	1
10	0,8	0,83	1	0,91	1	1

Таблица 4. Результаты работы сервисов по показателю Recall

Table 4. The results in terms of Recall

Текст №	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru
1	0,6	0,8	0,8	0,8	1	0,8
2	0,8	1	1	1	0,6	1
3	0,66	1	1	0,67	1	1
4	0,84	0,92	0,92	0,38	0,77	0,69
5	0,91	1	1	1	0,81	1
6	1	0,88	0,89	0,89	0,67	0,89
7	0,53	0,5	0,71	0,42	0,75	0,5
8	0,58	1	1	0,92	0,25	0,75
9	0,69	0,77	0,61	0,54	0,69	0,54
10	0,86	0,71	0,86	0,71	0,93	0,71

Таблица 5. Результаты работы сервисов по показателю F1- мера

Table 5. The results in terms of F1- мера

Текст №	TextRazor	Rosette Text Analytics	Eureka Engine	Texterra	Pullenti	NER-ru
1	0,67	0,61	0,89	0,80	1	0,89
2	0,73	0,91	0,91	1	0,6	0,91
3	0,44	0,86	1	0,57	1	0,67
4	0,87	0,96	0,96	0,52	0,77	0,75
5	0,74	0,88	1	0,96	0,81	0,96
6	0,67	0,73	0,94	0,84	0,63	0,89
7	0,55	0,58	0,83	0,6	0,72	0,58
8	0,37	0,85	1	0,96	0,2	0,78
9	0,54	0,69	0,76	0,64	0,69	0,70
10	0,83	0,77	0,92	0,8	0,96	0,83

Таблица 6. Средние значения показателей по результатам работы сервисов

Table 6. Average values of indicators based on service results

Сервисы	**Средние значения показателей**			
	A	**P**	**R**	**F1**
TextRazor	0,48	0,59	0,75	0,64
Rosette Text Analytics	0,66	0,74	0,86	0,78
Eureka Engine	**0,86**	**0,98**	**0,88**	**0,92**
Texterra	0,65	0,85	0,73	0,77
Pullenti	0,63	0,73	0,75	0,74
NER-ru	0,67	0,85	0,79	0,80

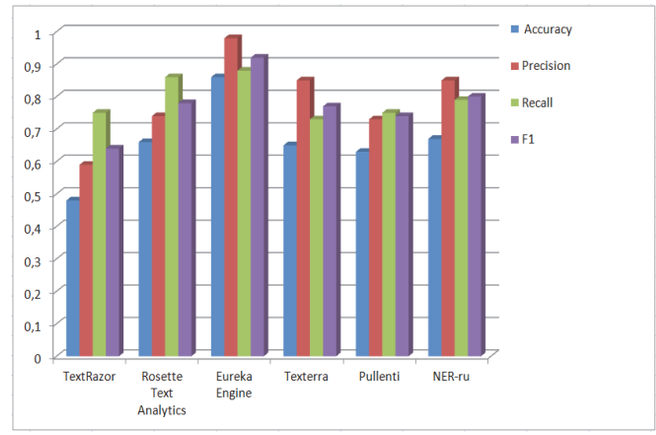

Рис. 1. Сравнительные показатели качества работы сервисов в задаче определения именованных сущностей

Fig. 1. Comparative indicators of the quality of work of services in the task of determining named entities

Результаты исследования работы сервисов по частеречной разметке

Производился анализ качества частеречной разметки сервисов: RosetteTextAnalytics, UDPipe. В связи с тем, что в TextRazor такая разметка для русского языка отсутствует, а в EurekaEngine она ограничена демо-версией и нестабильной работой частеречной разметки (разметка либо не работает, либо только для первых 10 слов текста), анализ разметки остальных сервисов планируется в будущем.

В первом тексте первого корпуса из коллекции FakeNews

из 157 слов ошибки в лемматизации были в 5 словах, частеречная разметка ошиблась в 5 словах, во втором тексте из 188 слов лемматизация была ошибочной для 6 слов, частеречная разметка для 9 слов, в третьем тексте из 180 слов ошибочная лемматизация имела место для 4 слов, частеречная разметка для 7 слов. Выяснилось, что сервис не воспринимает букву Ё. На втором корпусе, тексты в котором размечены вручную (OpenCorpora), RosetteTextAnalytics и UDPipe в качестве первоначальных результатов конкурируют между собой, поочередно занимая первое место.

Результаты исследования работы сервисов по тональной разметке

Был произведен анализ качества тональной разметки только одного сервиса EurekaEngine, в связи с тем, что это единственный сервис из списка анализируемых, который на это способен. На негативных текстах (10 шт) EurekaEngine оказался прав в 9 из 10 случаев. На позитивных текстах (10 шт) EurekaEngine оказался прав в 10 из 10 случаев.

Результаты исследования работы сервисов по категоризации

Таблица 7. Выделенные сервисами категории текстов
Table 7. Categories of text highlighted by services

Текст №	TextRazor	Eureka Engine	Rosette Text Analytics
1	Политика	Политика	Техника
2	Отдых и выходные	Шоу-бизнес и развлечения	Спорт
3	Вычислительные и информационные технологии	Экономика и бизнес	Закон
4	Транспорт, авиационные аварии и инциденты	Туризм и путешествия	Путешествия
5	Политика	Политика	Путешествия
6	Политика	Политика	Бизнес
7	Политика	Политика	Карьера
8	Религия и убеждения	Политика и бизнес	Религия
9	Экономика и бизнес	Экономика и бизнес	Бизнес
10	Политика	Политика	Закон

В результате TextRazor показал точность 10 из 10, Eureka Engine 7 из 10, Rosette Text Analytics 2 из 10.

Заключение

Использование SaaS имеет ряд преимуществ, позволяя разработчикам экономить аппаратные ресурсы и время на разработку и отладку своих продуктов. Проведенный анализ показал, что рассмотренные сервисы в целом успешно справляются с такими задачами обработки естественного языка, как частеречная разметка, анализ тональности, распознавание именованных сущностей, категоризации текстов и др., что позволяет использовать их как самостоятельно, так и как часть IT решения для обработки естественных текстов. Поскольку большая часть продуктов в этом сегменте в основном ориентирована на работу с английским языком, либо с переводом с других языков на английский, особое внимание было уделено тому, как они справляются с обработкой русского языка.

Анализ характеристик сервисов ОЕЯ был проведен на небольшом корпусе текстов, что, несомненно не позволяет ему в полной мере претендовать на объективность и полноту. Тем не менее, он позволяет сделать вывод о том, что при решении задач ОЕЯ на русском языке для распознавания именованных сущностей (таблица 6) и определения тональности текста лучше использовать облачный сервис EurekaEngine, а для частеречной разметки текста лучше использовать RosetteTextAnalytics, для категоризации текста лучше использовать TextRazor. Из сказанного выше видно, что ни один сервис не является абсолютным лидером, универсальным средством для решения всех задач ОЕЯ, у каждого есть свои сильные и слабые стороны. Поэтому, определенный сервис может быть выбран исходя из приведенных сравнительных характеристик и ценовой политики, наиболее приоритетных для пользователя, исходя из его задач и финансовых возможностей.

Дальнейшая работа в решении задачи сравнительного анализа сервисов предполагает использование более представительных корпусов и автоматизацию работы по оценке качественных показателей сервисов. Кроме того планируется рассмотреть качественные показатели доступных библиотек программ.

Благодарности

Работа выполнена в рамках темы «Разработка информационных технологий и систем для стимулирования устойчивого развития личности как одна из основ развития цифрового Казахстана» (программно-целевое финансирование МОН РК, ПЦФ BR05236839).

Acknowledgements

Work under the theme “Development of information technologies and systems to stimulate sustainable development of a person as one of the pillars of the development of digital Kazakhstan” (program-targeted financing of the Ministry of Education and Science of the Republic of Kazakhstan, PCF BR05236839).

Список использованных источников

[1] *Cambria E., White B.* Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article] // IEEE Computational Intelligence Magazine. 2014. Vol. 9, no. 2. Pp. 48-57. DOI: 10.1109/MCI.2014.2307227

[2] *Sreelekha S., Bhattacharyya P., Jha S.K., Malathi D.* A survey report on evolution of machine translation // International Journal of Control Theory and Applications. 2016. Vol. 9, no. 33. Pp. 233-240.

[3] *Höffner K., Walter S., Marx E., Usbeck R., Lehmann J., Ngomo A-CN.* Survey on Challenges of Question Answering in the Semantic Web // Semantic Web. 2017. Vol. 8, no. 6. Pp. 895-920. DOI: 10.3233/SW-160247

[4] *Qi W., Teney D., Wang P., Shen C., Dick A., van den Hengel A.* Visual question answering: A survey of methods and datasets // Computer Vision and Image Understanding. 2017. Vol. 163. Pp. 21-40. DOI: 10.1016/j.cviu.2017.05.001

[5] *Jurafsky D., Martin J.H.* Speech and Language Processing (2nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

2009. 950 p.

[6] *Niklaus C., Cetto M., Freitas A., Handschuh S.* A Survey on Open Information Extraction // CoRR. 2018. Vol. abs/1806.05599. URL: http://arxiv.org/abs/1806.05599 (дата обращения: 12.09.2018).

[7] *Deo A., Jayesh G., Shweta G.* A survey paper on information retrieval system // International Journal of Advanced Research in Computer Science. 2018. Vol. 9, no. 1. Pp. 778-781. DOI: 10.26483/ijarcs.v9i1.5505

[8] *Шокин Ю.И., Федотов А.М., Барахнин В.Б.* Проблемы поиска информации / Под ред. О.Л. Жижимова. Новосибирск: Изд. «Наука», 2010. 195 с. URL: https://elibrary.ru/item.asp?id=20126897 (дата обращения: 12.09.2018).

[9] *Campos R., Dias G., Jorge A.M., Jatowt A.* Survey of Temporal Information Retrieval and Related Applications // ACM Computing Surveys (CSUR). 2015. Vol. 47, issue 2. Article 15. 41 p. DOI: 10.1145/2619088

[10] *Purves R.S., Clough P., Jones C.B., Hall M.H., Murdock V.* Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text // Foundations and Trends® in Information Retrieval. 2018. Vol. 12, no. 2-3. Pp. 164-318. DOI: 10.1561/1500000034

[11] *Shiliang Sun, Chen Luo, Junyu Chen* A review of natural language processing techniques for opinion mining systems // Information Fusion. 2017. Vol. 36. Pp. 10-25. DOI: 10.1016/j.inffus.2016.10.004

[12] *Le Cun Y., Bengio Y., Hinton G.* Deep learning // Nature. 2015. Vol. 521. Pp. 436-444. DOI: 10.1038/nature14539

[13] *Hogenboom F.* et al. A survey of event extraction methods from text for decision support systems // Decision Support Systems. 2016. Vol. 85. Pp. 12-22. DOI: 10.1016/j.dss.2016.02.006

[14] *Potthast M., Hagen M., Stein B.* Author Obfuscation: Attacking the State of the Art in Authorship Verification // CLEF2016. Working Notes of CLEF 2016 – Conference and Labs of the Evaluation forum. Évora, Portugal, 5-8 September, 2016. CEUR Workshop Proceedings. Vol. 1609. Pp. 716-749. URL: http://ceur-ws.org/Vol-1609/16090716.pdf (дата обращения: 12.09.2018).

[15] *Hirschberg J., Manning C.D.* Advances in natural language processing // Science. 2015. Vol. 349, issue 6245. Pp. 261-266. DOI: 10.1126/science.aaa8685

[16] *Manning C.D., Schütze H.* Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, London, England, 1999. 680 p.

[17] *Goldberg Y.* A Primer on Neural Network Models for Natural Language Processing // Journal of Artificial Intelligence Research. 2016. Vol. 57, issue 1. Pp. 345-420.

[18] *Manning C.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., McClosky D.* The Stanford CoreNLP Natural Language Processing Toolkit // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations. Association for Computational Linguistics, Stroudsburg, PA, 2014. Pp. 55-60. DOI: 10.3115/v1/P14-5010

[19] *Dale R.* NLP meets the cloud // Natural Language Engineering. 2015. Vol. 21, issue 4. Pp. 653-659. DOI: 10.1017/S1351324915000200

[20] *Tablan V., Bontcheva K., Roberts I., Cunningham H., Dimitrov M.* AnnoMarket: An Open Cloud Platform for NLP // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2013. Pp. 19-24. URL: http://aclweb.org/anthology/P13-4004 (дата обращения: 12.09.2018).

[21] *Dale R.* Text Analytics APIs, Part 2: The Smaller Players // Natural Language Engineering. 2018. Vol. 24, issue 5. Pp. 797-803. DOI: 10.1017/S135132491800027X

[22] *Рубайло А.В., Косенко М.Ю.* Программные средства извлечения информации из текстов на естественном языке // Альманах современной науки и образования. 2016. № 12(114).С. 87-92. URL: https://elibrary.ru/item.asp?id=27710461 (дата обращения: 12.09.2018).

[23] *Марусова В.А.* Сравнительный анализ существующих систем для определения тональности текста // Материалы и методы инновационных исследований и разработок. Челябинск, 2016. Ч. 2. С. 66-68. URL: https://elibrary.ru/item.asp?id=27402188 (дата обращения: 12.09.2018).

[24] *Shavrina T.* Differential approach to webcorpus construction // Komp'juternaja lingvistika i intellektual'nye tehnologii. 2018. URL: https://elibrary.ru/item.asp?id=35716241 (дата обращения: 12.09.2018).

[25] *Muhamedyev R.* Machine learning methods: An overview // Computer Modelling & New Technologies. 2015. Vol. 19, no. 6. Pp. 14-29. URL: www.cmnt.lv/upload-files/ns_24rrt02_ReviewPaper.pdf (дата обращения: 12.09.2018).

Поступила 12.09.2018; принята в печать 20.10.2018;
опубликована онлайн 10.12.2018.

References

[1] Cambria E., White B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*. 2014; 9(2):48-57. DOI: 10.1109/MCI.2014.2307227

[2] Sreelekha S., Bhattacharyya P., Jha S.K., Malathi D. A survey report on evolution of machine translation. *International Journal of Control Theory and Applications*. 2016; 9(33):233-240.

[3] Höffner K., Walter S., Marx E., Usbeck R., Lehmann J., Ngomo A-CN. Survey on Challenges of Question Answering in the Semantic Web. *Semantic Web*. 2017; 8(6):895-920. DOI: 10.3233/SW-160247

[4] Qi W., Teney D., Wang P., Shen C., Dick A., van den Hengel A. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*. 2017; 163:21-40. DOI: 10.1016/j.cviu.2017.05.001

[5] Jurafsky D., Martin J.H. Speech and Language Processing (2nd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA. 2009. 950 p.

[6] Niklaus C., Cetto M., Freitas A., Handschuh S. A Survey on Open Information Extraction. *CoRR*. 2018. Vol. abs/1806.05599. Available at: http://arxiv.org/abs/1806.05599 (accessed 12.09.2018).

[7] Deo A., Jayesh G., Shweta G. A survey paper on information retrieval system. *International Journal of Advanced Research in Computer Science*. 2018; 9(1):778-781. DOI: 10.26483/ijarcs.v9i1.5505

[8] Shokin Yu.I., Fedotov A.M., Barakhnin V.B. Problems finding information. Novosibirsk: Nauka, 2010. 195 p. Available at: https://elibrary.ru/item.asp?id=20126897 (accessed 12.09.2018). (In Russian)

[9] Campos R., Dias G., Jorge A.M., Jatowt A. Survey of Temporal Information Retrieval and Related Applications. *ACM Computing Surveys (CSUR)*. 2015; 47(2):15. 41 p. DOI: 10.1145/2619088

[10] Purves R.S., Clough P., Jones C.B., Hall M.H., Murdock V. Geographic Information Retrieval: Progress and Challenges in Spatial Search of Text. *Foundations and Trends® in Information Retrieval*. 2018; 12(2-3):164-318. DOI: 10.1561/1500000034

[11] Shiliang Sun, Chen Luo, Junyu Chen A review of natural language processing techniques for opinion mining systems. *Information Fusion*. 2017; 36:10-25. DOI: 10.1016/j.inffus.2016.10.004

[12] Le Cun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015; 521:436-444. DOI: 10.1038/nature14539

[13] Hogenboom F. et al. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*. 2016; 85:12-22. DOI: 10.1016/j.dss.2016.02.006

[14] Potthast M., Hagen M., Stein B. Author Obfuscation: Attacking the State of the Art in Authorship Verification. CLEF2016. Working Notes of CLEF 2016 – Conference and Labs of the Evaluation forum. Évora, Portugal, 5-8 September, 2016. *CEUR Workshop Proceedings*. Vol. 1609, pp. 716-749. Available at: http://ceur-ws.org/Vol-1609/16090716.pdf (accessed 12.09.2018).

[15] Hirschberg J., Manning C.D. Advances in natural language processing. *Science*. 2015; 349(6245):261-266. DOI: 10.1126/science.aaa8685

[16] Manning C.D., Schütze H. Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, Massachusetts, London, England, 1999. 680 p.

[17] Goldberg Y. A Primer on Neural Network Models for Natural Language Processing. *Journal of Artificial Intelligence Research*. 2016; 57(1):345-420.

[18] Manning C.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, System Demonstrations*. Association for Computational Linguistics, Stroudsburg, PA, 2014, pp. 55-60. DOI: 10.3115/v1/P14-5010

[19] Dale R. NLP meets the cloud. *Natural Language Engineering*. 2015; 21(4):653-659. DOI: 10.1017/S1351324915000200

[20] Tablan V., Bontcheva K., Roberts I., Cunningham H., Dimitrov M. AnnoMarket: An Open Cloud Platform for NLP. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2013, pp. 19-24. Available at: http://aclweb.org/anthology/P13-4004 (accessed 12.09.2018).

[21] Dale R. Text Analytics APIs, Part 2: The Smaller Players. *Natural Language Engineering*. 2018; 24(5):797-803. DOI: 10.1017/S135132491800027X

[22] Rubailo A.V., Kosenko M.Yu. Software to extract information from texts in a natural language. *Almanac of modern science and education*. 2016; 12(114):87-92. Available at: https://elibrary.ru/item.asp?id=27710461 (accessed 12.09.2018).

[23] Marusova V.A. Comparative analysis of existing systems for determining the tonality of the text. *Materials and methods of innovative research and development*. Chelyabinsk, 2016; 2:66-68. Available at: https://elibrary.ru/item.asp?id=27402188 (accessed 12.09.2018).

[24] Shavrina T. Differential approach to webcorpus construction. *Komp'juternaja lingvistika i intellektual'nye tehnologii*. 2018. Available at: https://elibrary.ru/item.asp?id=35716241 (accessed 12.09.2018).

[25] Muhamedyev R. Machine learning methods: An overview. *Computer Modelling & New Technologies*. 2015; 19(6):14-29. Available at: www.cmnt.lv/upload-files/ns_24rrt02_ReviewPaper.pdf (accessed 12.09.2018).

Submitted 12.09.2018; revised 20.10.2018;
published online 10.12.2018.

About the authors:

Ravil I. Mukhamediev, Doctor of Engineering, Professor, Chief Researcher, Institute of Information and Computational Technologies MES RK (125 Pushkin Str., Almaty 050010, Republic of Kazakhstan); Professor, Department of Software Engineering, Institute of Information and Telecommunication Technologies, Kazakh National Research Technical University named after K.I. Satpaev (22a Satpaev Str., Almaty 050013, Republic of Kazakhstan); Professor, Dr.Sc., ISMA University (1 Lomonosova Str., Bldg. 6, Riga LV-1019, Latvia), ORCID: http://orcid.org/0000-0002-3727-043X, ravil.muhamedyev@gmail.com

Adilkhan Symagulov, Bachelor of Natural Sciences, Software Engineer, Institute of Information and Computational Technologies MES RK (125 Pushkin Str., Almaty 050010, Republic of Kazakhstan); undergraduate, Kazakh National Research Technical University named after K.I. Satpaev (22a Satpaev Str., Almaty 050013, Republic of Kazakhstan), ORCID: http://orcid.org/0000-0001-9974-3215, asmogulove00@gmail.com

Yan I. Kuchin, Master, Researcher, Institute of Information and Computational Technologies MES RK (125 Pushkin Str., Almaty 050010, Republic of Kazakhstan), ORCID: http://orcid.org/0000-0002-5271-9071, ykuchin@mail.ru

Sabina Abdullayeva, Bachelor of Mathematical Methods in Economics, Laboratory Assistant, Institute of Information and Computational Technologies MES RK (125 Pushkin Str., Almaty 050010, Republic of Kazakhstan), ORCID: http://orcid.org/0000-0002-2212-0500, Sabina.abdullaeva23@gmail.com

Farida N. Abdoldina, Candidate of Technical Sciences, deputy director of the Institute of Information and Telecommunication Technologies, Kazakh National Research Technical University named after K.I. Satpaev (22a Satpaev Str., Almaty 050013, Republic of Kazakhstan), ORCID: http://orcid.org/0000-0003-1816-6343, farida_mail@mail.ru