Systems and Technologies Department, PhD in Technical Science. Samara State Technical University, 244, Molodogvardeyskaya Street, Samara, 443100, Russian Federation; Associated Professor of Technologies Department. Tel. +7 927 706-61-21. E-mail: s.palmov@psuti.ru

Shatalov Nikita Vladimirovich, Povolzhskiy State University of Telecommunications and Informatics, 77, Moscovskoye shosse, Samara, 443090, Russian Federation; Student of Information Systems and Technologies Department. Tel. +7 999 170-27-28. E-mail: nickit.schatalow@yandex.ru

References

1. What is Vaex? URL: https://vaex.readthedocs.io/en/latest/index.html (accessed: 15.04.2024).
2. Dask – Dask documentation. URL: https://docs.dask.org/en/stable/ (accessed: 15.04.2024).
3. GitHub – dask/dask: Parallel computing with task scheduling. URL: https://github.com/dask/dask (accessed: 16.04.2024).
4. NumPy. URL: https://numpy.org/ (accessed: 16.04.2024).
5. GitHub – vaexio/vaex. URL: https://github.com/vaexio/vaex (accessed: 17.04.2024).
6. Dask vs Vaex – a qualitative comparison. URL: https://vaex.io/blog/dask-vs-vaex-a-qualitative-comparison (accessed: 17.04.2024).
7. How to use HDF5 files in Pytho. URL: https://habr.com/ru/companies/otus/articles/416309/ (accessed: 17.04.2024). (In Russ.)
8. 52 datasets for training projects. URL: https://habr.com/ru/companies/edison/articles/480408/ (accessed: 18.04.2024). (In Russ.)
9. Vaex and Dask: when Pandas cannot process big data. URL: https://python-school.ru/blog/analiz-dannyh/vaex-vs-dask/ (accessed: 18.04.2024). (In Russ.)
10. Using the Vaex library for processing large amounts of data. URL: https://newtechaudit.ru/ispolzovanie-biblioteki-vaex-dlya-obrabotki-bolshih-obyomov-dannyh/ (accessed: 19.04.2024). (In Russ.)
11. Data analysis using the Dask library. URL: https://habr.com/ru/companies/otus/articles/759552/ (accessed: 19.04.2024). (In Russ.)
12. Gruzdev A.V., Heidt M. *Studying Pandas*. Transl. From English by A.V. Gruzdev. Moskow: DMK, 2019, 682 p. (In Russ.)
13. Ues M. *Python and Data Analysis. Primary Data Processing Using Pandas, Numpy and Jupiter*. Transl. From English by A.A. Slinkin. 3nd ed. Moscow: DMK, 536 p. (In Russ.)
14. Vasiliev Yu.A. *Python for Data Science.* Saint Petersburg: Piter, 272 p. (In Russ.)

Received 23.04.2024

УДК 004.89

РАЗРАБОТКА ИНСТРУМЕНТА ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ РЕШЕНИЯ ПРИКЛАДНОЙ ЗАДАЧИ ИЗВЛЕЧЕНИЯ СТАТИСТИЧЕСКИХ ДАННЫХ ИЗ ТЕКСТА

Захарова О.И., Бедняк С.Г.

Поволжский государственный университет телекоммуникаций и информатики, Самара, РФ

E-mail: o.zaharova@psuti.ru

Текстовая аналитика используется для изучения текстового содержимого и получения новых переменных из необработанного текста, которые можно использовать в качестве входных данных для моделей прогнозирования или других статистических методов, в том числе при решении фундаментальных задач. Цель исследования: проанализировать алгоритмы машинного обучения, практические наработки в этой области и разработать интегрируемый программный инструмент обработки текста, используя структуру алгоритма, на основе библиотек BasicStats, ReadabilityStats, SovChLit, позволяющий извлекать статистику из текстов большого объема на русском языке. Реализован метод извлечения статистических данных из необработанных текстов больших объемов на основе машинного обучения и обработки естественного языка на языке Python, с возможностью встраивания в другие проекты. Разработан программный инструмент, использующий функционал адаптированной для русского языка

библиотеки textary, который позволяет работать как с текстами, так и с Doc-объектами, подготовленными с помощью библиотеки spaCY. Для проведения исследования были задействованы реальные текстовые данные, собранные с информационно-новостного портала по Самарской области «63.ru» (в рамках реализации концептуального проекта «Ферма данных» научно-исследовательской лаборатории искусственного интеллекта). Разработанный программный инструмент извлечения статистических данных из текста позволяет анализировать большие объемы текстовых данных и извлекать из них полезную информацию. Его можно интегрировать в другие программные решения, как один из связующих модулей в цепи оптимизации кода для программ по обработке текстовых данных.

Ключевые слова: *Natural Language Processing, алгоритм обработки естественного языка, обработка текста, извлечение статистических данных, машинное обучение, Python*

Введение

Обработка текстов на естественном языке (Natural Language Processing, NLP) – наиболее заметно развивающиеся направление, за которым стоит несколько десятков лет фундаментальных исследований и плодотворной работы в области лингвистики, математики и информатики. Термин относят к области искусственного интеллекта, в основе идеи лежит процесс интерпретации и преобразования необработанного письменного текста в форму, понятную машине [1; 2].

NLP в последнее время активно применяется в повседневной жизни человека и становится все более важным направлением по мере того, как языковые технологии используются в различных областях. Яркими примерами разработок на основе NLP, с которыми многие сталкиваются в повседневной жизни, являются голосовые системы GPS, цифровые помощники и т.д. Но NLP также играет все более важную роль в корпоративных решениях, помогающих упростить бизнес-операции, повысить производительность сотрудников и упростить критически важные бизнес-процессы.

Актуальность NLP и развитие одного из важнейших ее этапов, семантического анализа [3], подтверждается тем, что ряд мировых компаний создают собственные экосистемы и сервисы анализа текста. Стоит отметить и ряд опубликованных российских проектов [4; 5]:

– языковая модель RuBERT от DeepPavlov [21; 22];

– бенчмарк для русского языка RussianSuperGlue.

Кроме того, такие LLM (Large Language Model) как ChatGPT, подчеркивают важность и своевременность проводимых исследований и программных разработок в данной области [6].

Применение рассматриваемых методов и технологий для решения прикладных задач в области NLP [23], в том числе извлечение статистических показателей по текстовым данным, может быть использовано в программных разработках по обработке текстов. Например, их можно применять в анализе тональности текста [12; 13; 16; 17], определении частоты использования определенных слов или фраз, выявлении ключевых тем или идей, представленных в тексте, а также многими другими способами [14; 15]. Ряд таких проектов реализуется на базе научно-исследовательской лаборатории искусственного интеллекта (НИЛ ИИ) Поволжского государственного университета телекоммуникаций и информатики (ПГУТИ). Статистические показатели, собираемые разрабатываемым инструментом, будут применяться при работе с большими объемами текста, такими как новостные, научные статьи и социальные медиа-посты. Они позволят быстро получить представление о содержании текста и выделить наиболее важные аспекты.

В связи с этим актуальной является задача по программной реализации алгоритма, реализующего описываемые функциональные возможности, особенно на основе русскоязычных текстов, за счет использования современных библиотек. В рамках разработки программного инструмента на основе рассматриваемого алгоритма была взята за основу библиотека textacy (библиотека Python для выполнения различных задач NLP, построенная на высокопроизводительной библиотеке spaCy).

В свою очередь, spaCy – это библиотека для расширенной обработки в Python и Cython. Она основана на самых последних исследованиях и с самого начала разрабатывалась для использования в реальных продуктах. SpaCy предлагает предварительно обученные конвейеры, которые позволяют работать с текстом на более чем 60 языках. SpaCy также поддерживает многозадачное обучение с использованием предварительно обученных преобразователей, таких как BERT (Bidirectional Encoder Representations from Transformers).

Алгоритм извлечения статистических данных из текстов

Как уже говорилось выше – процесс обработки естественного языка [19] состоит из двух основных этапов: понимания естественного языка и генерации естественного языка.

NLU (Natural Language Understanding) пытается понять смысл данного текста. Характер и структура каждого слова внутри текста должны быть из-

вестны для NLU. Для понимания структуры NLU пытается разрешить следующую двусмысленность, присутствующую в естественном языке [7]:

– лексическая неоднозначность;
– синтаксическая неоднозначность;
– семантическая неоднозначность [11];
– анафорическая двусмысленность.

Затем смысл каждого слова понимается с помощью лексикона (лексики) и набора грамматических правил.

NLG (Natural Language Generation) – это процесс автоматического создания текстовых данных из структурированных данных в удобочитаемом формате со значимыми фразами и предложениями [1; 2]. С проблемой генерации естественного языка трудно справиться. Процесс реализации чаще всего включает в себя три этапа:

– планирование текста – выполняется упорядочение основного контента в структурированные данные;

– планирование предложений. Предложения объединяются со структурированными данными для представления потока информации;

– реализация. В конечном итоге для представления текста создаются грамматически правильные предложения.

В ходе нашей работы мы использовали несколько методов для разработки алгоритма по обработке текста:

1. Простейшие метрики [8].

Обработка естественного языка обычно означает обработку текста или текстовой информации (аудио, видео). Важным этапом в этом процессе является преобразование разных слов и словоформ в одну речевую форму. Кроме того, часто нужно измерить, насколько похожи или различны строки. Обычно в этом случае используем различные метрики, показывающие разницу между словами.

Одной из простых и в то же время широко используемых метрик является расстояние редактирования – алгоритм оценки схожести двух строковых значений (слово, форма слова, состав слова), путем сравнения минимального количества операций по преобразованию одного значения в другое.

2. Векторизация [8].

Векторизация – это процедура преобразования слов (текстовой информации) в цифры для извлечения текстовых атрибутов (признаков) и дальнейшего использования алгоритмов машинного обучения (NLP). Другими словами, векторизация текста представляет собой преобразование текста в числовые векторы.

Была подготовлена схема поэтапного выполнения блоков алгоритма (рисунок 1), использующего библиотеки, в работе по процессу извлечения объектов выстроенного алгоритма [18].

Более полно рассмотренную схему взаимодействия было решено реализовать по основным этапам обработки текстовых данных в формате схематичного алгоритма на языке Python для решения прикладной задачи (рисунок 2).

В качестве основного инструмента для решения рассмотренных задач был выбран язык Python, среда разработки Google Colab и набор инструментов для обработки естественного языка Natural Language Toolkit (NLTK).

NLTK – это платформа для создания проектов Python, популярная благодаря своим массивным корпусам, обилию библиотек и подробной документации [7].

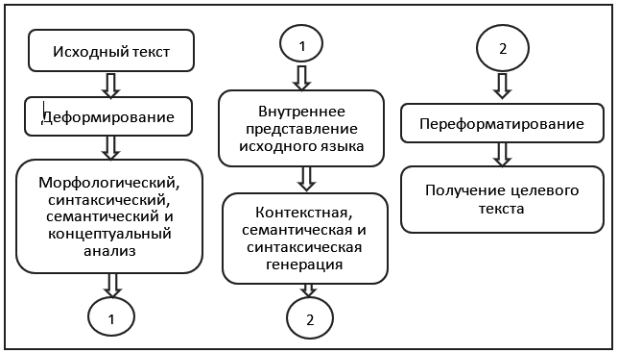

Рисунок 1. Основные этапы обработки текста, реализованные в разрабатываемом алгоритме

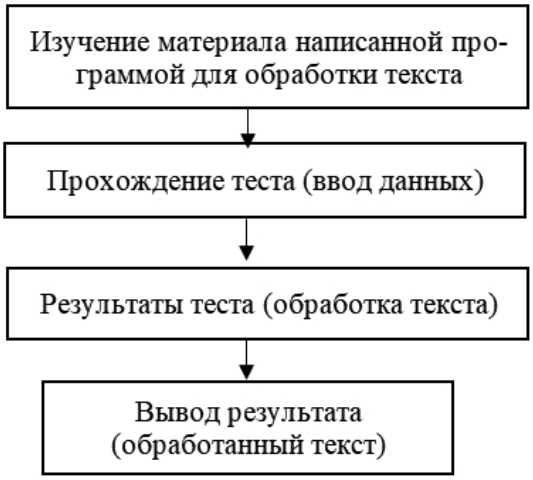

Рисунок 2. Этапы обработки текста для извлечения статистических данных

Google Colab представляет собой бесплатный онлайн-сервис для работы с машинным обучением и анализа данных, предлагаемый корпорацией Google, и предоставляет пользователям доступ к мощным вычислительным ресурсам, таким как GPU-ускорение (графические ускорители) и TPU (тензорные процессоры), что позволяет ускорить процесс обработки текстовых данных и обеспе-

чивает получение более точных результатов. Это делает его идеальным инструментом для разработки и тестирования новых алгоритмов и методов обработки текстовых данных.

Исходя из разработанной схемы поэтапного выполнения модулей программного инструмента (рисунок 1), использующего библиотеки в работе по процессу извлечения объектов [20] выстроенного алгоритма, была реализована программа, написанная на яыке Python.

Процесс извлечения объектов в рамках реализации алгоритма работы метода NLP для сбора статистики из текстов

Для проведения исследования были задействованы реальные данные, собранные с информационно-новостного портала по Самарской области «https://63.ru/» в рамках реализации концептуального проекта «Ферма данных» НИЛ ИИ ПГУТИ, где ежедневно публикуются новости, обсуждаются проблемы региона и сообщается о происшествиях. Собранная структура данных представлена на рисунке 3.

Вид собираемых данных в формате полнотекстовых записей (тексты статей), а также комментарии для этих тестов, для дальнейшей обработки разрабатываемым инструментом наглядно показывают их неоднородность и различную тематическую направленность (рисунок 4).

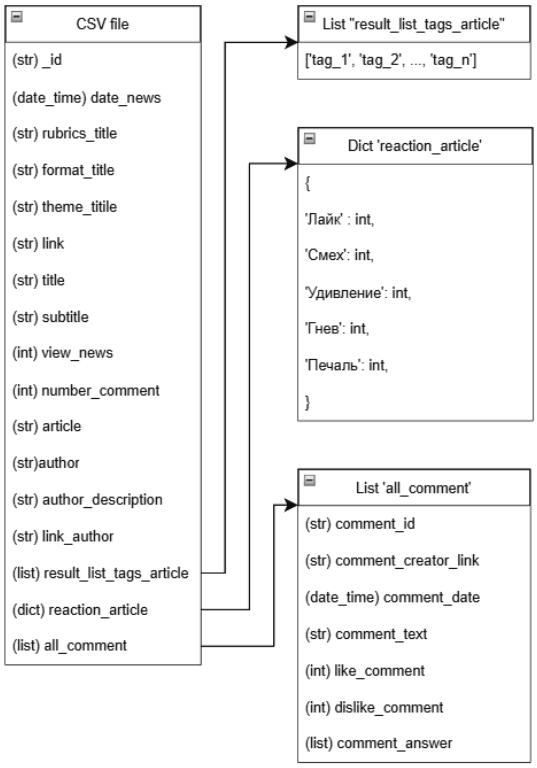

Рисунок 3. Структура данных, полученных с портала «https://63.ru/»

Полные статьи

[3]:
```
text_1 = df_1['article'][0]
print(text_1)
```

```
В «Трамвайно-троллейбусном управлении» Самары придумали, как модернизировать старые трамваи. Предприятие планирует начать реализаци
ю программы капитально-восстановительного ремонта. Об этом в интервью 63.RU рассказал директор ТТУ Михаил Ефремов. — Что она из себ
я представляет? «Татру» разбирают до железного «скелета», на который «навешивают» новую проводку, стекло, кресла. Это позволяет про
длить срок службы трамвая, чтобы он смог проехать еще 300 тысяч километров. И сейчас этот вопрос мы прорабатываем вместе с департам
ентом транспорта. Уже нашли компанию в Екатеринбурге, которая занимается модернизацией трамваев. Они сотрудничают с местным горэлек
тротранспортом. Правда, их специалисты работали с другими вагонами от другого производителя. Но мы уверены, что и с «Татрами» справ
ятся. Ожидаем, что к концу осени у нас уже появится первый модернизированный трамвай, — пояснил Михаил Ефремов. Такая модернизация
— тоже удовольствие не из дешевых. По предварительным подсчетам, один вагон обойдется в 19 миллионов рублей. Но если всё пойдет по
плану, то реализация программы начнется уже в этом году. По сути, это будет новый трамвай, с современной системой безопасности, кон
диционирования, мягкими сиденьями. Можно даже поменять внешний вид трамвая. — Предполагается, что модернизация станет совместным пр
оектом с участием Северного и Городского трамвайного депо и коллег из Екатеринбурга. Если всё получится, то будет польза и предприя
тию, и всему городу. Можно будет по 20 трамваев в год ремонтировать. Это хорошая цифра, колоссальный объем, — резюмировал директор
ТТУ. Он уточнил, что если программа будет длиться минимум 5 лет, то получилось бы отремонтировать 100 трамваев.
```

[4]:
```
text_2 = df_1['article'][1]
print(text_2)
```

```
Бездомные псы, на первый взгляд, бывают милыми. Но от них можно ожидать всего, чего угодно В Кошкинский районный суд подали иск о в
зыскании морального и материального вреда с владелицы собаки. Об этом рассказали в прокуратуре Самарской области. Собака напала на
13-летнего мальчика в феврале 2022 года. Пес укусил подростка за кисть руки. Мальчику сделали цикл уколов от бешенства. Семья мальч
ика потратила деньги на лекарства, а сам он испытал нравственные страдания. — Прокуратура во время проверки выяснила, что нападение
собаки на ребенка случилось по вине собственницы животного — жительницы села Русская Васильевка, — говорится в сообщении облпрокура
туры. Теперь с собственницы пса требуют компенсацию в 100 000 рублей. Нападение собак на людей, к сожалению, очень часто встречающе
еся явление в Самарской области. Недавно стая собак набросилась на двух школьниц в Новокуйбышевске. Одной из девочек пес прокусил б
едро. Еще один случай произошел в Самаре: здесь бездомные собаки напали на мальчика. Самую оперативную информацию о жизни Самары и
области мы публикуем в нашем телеграм-канале 63.RU. А в нашей группе во «ВКонтакте» вы можете предложить свои новости, истории, фот
ографии и видео.
```

[5]:
```
text_3 = df_1['article'][10]
print(text_3)
```

```
Помните, как лет 20-30 назад бразильские сериалы заполонили телеканалы? Каждый вечер мы стремились к экранам, чтобы вникать в увлек
ательные перипетии загадочной латиноамериканской жизни. Мы решили вернуть вам эти беззаботные времена, а заодно проверить, насколь
о хорошо вы помните всё, что смотрели. Справятся только те, кто смотрел «Клон» — но его одного, конечно же, не хватит!
```

Рисунок 4. Вид собираемых данных в формате полнотекстовых записей

При этом для решения ряда фундаментальных и практических задач и сбора дополнительных статистических данных, используемых в различных математических моделях по проектам лаборатории, возникла необходимость разработки готового интегрируемого инструмента извлечения статистических данных из необработанных текстов.

В результате проведенного анализа предметной области, тестирования наработок в этом отношении, а также современных средств машинного обучения, был разработан программный инструмент, основанный на описываемом выше функционале рассматриваемой библиотеки textary [7; 9].

Ниже приведен фрагмент кода, реализующий данный алгоритм:

```
import re
from nltk.corpus import stopwords
from ruts import SentsExtractor, WordsExtractor
import nltk
nltk.download('stopwords')
text = «Каждый день, каждый вечер, я выходил на улицу. Улица была тихая, улица была спокойная. Шаг за шагом, шаг за шагом, я шел вперед. Вдыхая свежий воздух, вдыхая ароматы весны. С каждым шагом, с каждым вдохом, я чувствовал себя свободным.»
se = SentsExtractor(tokenizer=re.compile(r', '))
se.extract(text)
we = WordsExtractor(use_lexemes=True, stopwords=stopwords.words('russian'), filter_nums=True, ngram_range=(1, 2))
we.extract(text)
we.get_most_common(3)
```

На рисунке 5 представлен результат работы алгоритма, использующего данную библиотеку.

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[('шаг', 5), ('каждый', 4), ('улица', 3)]
```

Рисунок 5. Результат работы алгоритма по обнаружению повторяющихся слов

Библиотека BasicStats, также применявшаяся в разработке программного инструмента, позволяет извлекать из текста различные статистические показатели.

Python statistics – это внутренняя библиотека Python для описательной статистики. Ниже приведен фрагмент кода, реализующий алгоритм описательной статистики:

```
from ruts import BasicStats
text = «Он смотрел вдаль, словно ища что-то, что ускользнуло из его памяти, но так и не нашел.»
bs = BasicStats(text)
bs.get_stats()
bs.print_stats()
```

На рисунке 6 представлен результат работы библиотеки BasicStats.

Статистика	Значение
Предложения	1
Слова	16
Уникальные слова	16
Длинные слова	5
Сложные слова	1
Простые слова	15
Односложные слова	8
Многосложные слова	8
Символы	86
Буквы	66
Пробелы	15
Слоги	27
Знаки препинания	5

Рисунок 6. Вывод результата с использованием библиотеки BasicStats

Примененная библиотека ReadabilityStats в разработанном программном инструменте (рисунок 6) предоставляет возможность вычисления метрик для текстовых данных.

Коэффициенты метрик для русского языка были взяты из работы исследователей проекта Plain Russian Language [10], которые получили их на основе специально подобранных текстов с предварительными возрастными пометками.

Ниже приведем фрагмент кода программного средства, реализующий алгоритм обработки текста по метрикам:

```
from ruts import ReadabilityStats
text = «После заката, когда небо окуталось сумерками, он отправлялся на прогулку, вдыхая свежий воздух и наслаждаясь тишиной и спокойствием окружающего мира.»
rs = ReadabilityStats(text)
rs.get_stats()
rs.print_stats()
```

После запуска кода с использованием библиотеки ReadabilityStats видим результат обработки текста по метрикам на рисунке 7, которые были описаны ранее.

Метрика	Значение
Тест Флеша-Кинкайда	12.92
Индекс удобочитаемости Флеша	18.56
Индекс Колман-Лиау	12.57
Индекс SMOG	21.71
Автоматический индекс удобочитаемости	14.01
Индекс удобочитаемости LIX	80.00

Рисунок 7. Вывод результата использования библиотеки ReadabilityStats в работе структуры алгоритма

Библиотека SovChLi, применяемая в алгоритме (рисунок 8), обеспечивает возможность работы с несколькими предварительно обработанными текстовыми наборами данных.

Далее приведем фрагмент кода с использованием данной библиотеки на практике:

```
from ruts.datasets import SovChLit
svc = SovChLit()
svc.download()
sc = SovChLit()
sc.info
for i in sc.get_records(max_len=100, category='Лето', limit=1):
    print(i)
for i in sc.get_texts(text_type='Рассказ', limit=1):
    print(i)
```

Таблица 1. Перечень библиотек, реализованных в разработанном программном инструменте

	Основная работа библиотеки
BasicStats	Вычисляет обзор основных статистических значений
ReadabilityStats	Для вычисления основных метрик удобочитаемости текста. В качестве источника данных может использоваться как непосредственно текст, так и объект класса Doc библиотеки spaCy
SovChLit	Выполняет загрузку набора данных из сети и извлечение файлов

В таблице 1 сформулирован основной функционал применяемых в разработанном программном средстве и описанных выше библиотек.

Заключение

Анализ существующих наработок в области обработки текстовых данных, в частности на базе русскоязычных текстов, показал, что обработка текстовых данных – это сложная задача, требующая большого объема данных и вычислительных ресурсов. Кроме того, результаты могут быть не всегда точными, особенно если тексты содержат неоднозначности или сложные конструкции. В рамках реализации ряда проектов на базе лаборатории искусственного интеллекта извлечение статистических данных из необработанных текстов является важным инструментом для анализа больших объемов информации и может быть использовано в различных областях для решения фундаментальных задач и применения этих данных в качестве параметров при математическом моделировании, что создает необходимость разработки интегрируемого в такие проекты программного инструмента.

Разработан программный инструмент, основанный на алгоритме извлечения статистических данных из текстов. Для этого была написана программа на языке Python, использующая библиотеки обработки текста NLTK и SpaCy, а ее работоспособность протестирована в среде Google Colab на различных тематиках текстов, собираемых в рамках текущих проектов НИЛ ИИ ПГУТИ.

Разработанное программное средство позволяет осуществлять извлечение объектов, подсчет базовой статистики, показателей удобочитаемости и метрики лексического разнообразия текстовыми данными в процессе сбора и обработки текстовых данных.

Проведенное исследование позволяет обозначить контуры дальнейшего направления работы с большими массивами текстов. В соответствии с проведенной работой по исследованию тематики обработки естественного языка, разработке соответствующего программного инструмента на основе машинного обучения для такой обработки и демонстрации его работы на примерах по извлечению статистики из текстовых данных, собираемых в проектах лаборатории, можно сделать вывод о практической значимости предложенного решения, связанной с его возможным применением, как одного из связующих звеньев в цепи оптимизации кода для программ по обработке текстов. Таким образом, можно выделить возможные точки роста и новые направления исследований.

```
Файл /usr/local/lib/python3.10/dist-packages/ruts_data/texts/sov_chrest_lit.tar.xz уже загружен
('В саду у Мичурина даже рябина стала хорошим плодовым деревом: ягоды на ней '
 'родились крупные, сочные и сладкие. А низенькие вишнёвые деревья летом были '
 'густо обсыпаны ягодами. На яблонях висели душистые, румяные яблоки, да такие '
 'крупные, что одно яблоко едва держишь в руках.\n'
 'Мичурин много и упорно работал. Советская власть помогала ему в работе.\n'
 'Мичурин вырастил новые, чудесные сорта яблонь, груш, вишен и винограда. '
 'Теперь во всех колхозных садах растут мичуринские сорта яблонь и груш, слив '
 'и вишен.')
```

Рисунок 8. Использование библиотеки SovChLit в работе алгоритма

Благодарность

Авторы выражают благодарность профессору, к.ф.-м.н., заведующему НИЛ ИИ ПГУТИ, Левашкину Сергею Павловичу за помощь в подготовке статьи и ценные комментарии.

Литература

1. Захарова О.И. Разработка системы анализа и обработки текстовых данных // Проблемы техники и технологий телекоммуникаций (ПТиТТ-2023): материалы XXV Международной научно-технической конференции. Казань: КНИТУ-КАИ, 2023. С. 261–262.
2. Кулешов С.В., Зайцева А.А., Левашкин С.П. Технологии и принципы сбора и обработки неструктурированных распределенных данных с учетом современных особенностей предоставления медиа-контента // Информатизация и связь. 2020. № 5. С. 22–28. DOI: 10.34219/2078-8320-2020-11-5-22-28
3. Захарова О.И. Семантический анализ и синтез текстовых данных. Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2023. № 4. С. 182–208. DOI: 10.17308/sait/1995-5499/2023/4/182-208
4. Smetanin S., Komarov M. Deep transfer learning baselines for sentiment analysis in Russian // Information Processing & Management. 2021. Vol. 58, no. 3. P. 102484. DOI:10.1016/j.ipm.2020.102484. URL: https://www.sci-hub.ru/10.1016/j.ipm.2020.102484 (дата обращения: 28.06.2024).
5. Шаврина Т.О. О методах компьютерной лингвистики в оценке систем искусственного интеллекта // Вопросы языкознания. 2021. № 6. С. 117–138. DOI: 10.31857/0373-658X.2021.6.117-138
6. Лещинская Н.М., Колесник М.А. Внедрение технологий искусственного интеллекта в России // Социология искусственного интеллекта. 2023. Т. 4, № 2. С. 63–72.
7. Comparing automated text classification methods / J. Hartmann [et al.] // International Journals of Research Marketing. 2019. Vol. 36, no. 1. P. 20–36. DOI:10.1016/j.ijresmar.2018.09.009. URL: https://www.sci-hub.ru/10.1016/j.ijresmar.2018.09.009 (дата обращения: 20.07.2024).
8. A robustly optimized BERT pretraining approach / Y. Liu [et al.]. URL: https://arxiv.org/pdf/1907.11692.pdf (дата обращения: 28.07.2024).
9. Захарова О.И., Левашкин С.П., Иванов К.Н. Современные библиотеки Python для сбора данных из интернета // Проблемы техники и технологий телекоммуникаций (ПТиТТ-2020): материалы XXII Международной научно-технической конференции. Самара: ПГУТИ, 2020. С. 316–317.
10. A novel machine learning approach for scene text extraction / G.J. Ansari [et al.] // Future Generation Computer Systems. 2018. Vol. 87. P. 328–340. DOI: 10.1016/J.FUTURE.2018.04.074
11. Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning / H. Kim [et al.] // Neurocomputing. 2018. Vol. 315. P. 128–134. DOI: 10.1016/J.NEUCOM.2018.07.002
12. Deep learning for affective computing: Text-based emotion recognition in decision support / B. Kratzwald [et al.] // Decision Support Systems. 2018. Vol. 115. P. 24–35. DOI: 10.1016/J.DSS.2018.09.002
13. Web opinion mining and sentimental analysis / E.M. Taylor [et al.] // Advanced Techniques in Web Intelligence-2. P. 105–126. DOI: 10.1007/978-3-642- 33326-2_5
14. A hybrid model of sentimental entity recognition on mobile social media / Z. Wang [et al.] // EURASIP Journal on Wireless Communications and Networking. DOI: 10.1186/s13638-016-0745-7. URL: https://sci-hub.ru/10.1186/s13638-016-0745-7 (дата обращения: 25.08.2024).
15. Altınel B., Ganiz M.C. Semantic text classification: A survey of past and recent advances // Information Processing & Managemen. 2018. Vol. 54, no. 6. P. 1129–1153. DOI: 10.1016/J.IPM.2018.08.001
16. Understanding emotions in text using deep learning and big data / A. Chatterjee [et al.] // Computers in Human Behavior. 2019. Vol. 93. P. 309–317. DOI: 10.1016/J.CHB.2018.12.029
17. Sentiment analysis of tweet data / S.M. Mazharul [et al.] // Hoque Chowdhury. URL: https://www.researchgate.net/publication/324965434_SENTIMENT_ANALYSIS_OF_TWEET_DATA (дата обращения: 27.4.2024).
18. A brief survey of text mining: classification, clustering and extraction techniques / M. Allahyari [et al.]. URL: https://www.researchgate.net/publication/318336890_A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques (дата обращения: 30.08.2024).
19. A survey on recent approaches for natural language processing in low-resource scenarios / M.A. Hedderich [et al.] // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021. P. 2545–2568.

20. Popovski G., Seljak B.K., Eftimov T. A survey of named-entity recognition methods for food information extraction // IEEE Access. 2020. Vol. 8. P. 31586–31594. DOI: 10.1109/ACCESS.2020.2973502
21. Bert: Pre-training of deep bidirectional transformers for language understanding / J. Devlin [et al.]. URL: https://arxiv.org/pdf/1810.04805 (дата обращения: 29.08.2024).
22. Spatish pre-trained BERT model and evaluation data / J. Canete [et al.] // Accepted as a workshop paper at PML4DC (ICLR). URL: https://www.researchgate.net/publication/372962444_Spanish_Pre-trained_BERT_Model_and_Evaluation_Data (дата обращения: 20.08.2024).
23. Иванов К.Н., Захарова О.И. Обработка естественного языка. Применение языковых моделей // Актуальные проблемы информатики, радиотехники и связи: материалы XXX Российской научно-технической конференции. Самара: ПГУТИ, 2023. С. 155–156.

Получено 05.08.2024

Захарова Оксана Игоревна, к.т.н., доцент, доцент кафедры информационных систем и технологий (ИСТ), заместитель заведующего научно-исследовательской лабораторией искусственного интеллекта (НИЛ ИИ) Поволжского государственного университета телекоммуникаций и информатики (ПГУТИ). 443090, Российская Федерация, г. Самара, Московское шоссе, 77. Тел. +7 906 343-25-21. E-mail: o.zaharova@psuti.ru

Бедняк Светлана Геннадьевна, к.п.н., доцент, доцент кафедры ИСТ ПГУТИ. 443090, Российская Федерация, г. Самара, Московское шоссе, 77. Тел. +7 903 308-29-88. E-mail: s.bednyak@psuti.ru

DEVELOPMENT OF A NATURAL LANGUAGE PROCESSING TOOL FOR SOLVING THE APPLICATION PROBLEM OF EXTRACTING STATISTICAL DATA FROM TEXT

Zakharova O.I., Bednyak S.G.
Povolzhskiy State University of Telecommunications and Informatics, Samara, Russian Federation
E-mail: o.zaharova@psuti.ru

Text analytics is used to explore textual content and obtain new variables from raw text, which can be used as input data for forecasting models or other statistical methods, including for solving fun-damental problems. The purpose of the research: to analyze machine learning algorithms, practical developments in this field and to develop an integrated software instrument for text processing, us-ing the structure of the algorithm, based on the BasicStats, ReadabilityStats, SovChLit libraries, al-lowing to extract statistics from raw texts of large volumes in Russian. A method of extracting sta-tistical data from raw texts of large volumes based on machine learning and natural language pro-cessing in Python has been implemented, with the possibility of embedding it into other projects. A software instrument that use the functionality of textary library adapted for Russian language was developed, which allows to work with both texts and Doc-objects generated with spaCY library. The study was conducted using real text data collected from the information and news portal for the Samara region «63.ru» (in the context of the implementation of the conceptual project «Data Farm» by the artificial intelligence research laboratory). The developed software for extracting statistical data from text allows analyzing large volumes of text data and extracting useful information from them. It can be integrated into other software solutions as one of the linking modules in the of code optimization chain for text data processing programs.

Keywords: *Natural Language Processing, natural language processing algorithm, text processing, statistical extraction, machine learning, Python*

DOI: 10.18469/ikt.2024.22.1.13

Zakharova Oksana Igorevna, Povolzhskiy State University of Telecommunications and Informatics, 77, Moscovskoye shosse, Samara, 443090, Russian Federation; Associated Professor of Information Systems and Technologies Department, Deputy Head of the Research Laboratory of Artificial Intelligence, PhD in Technical Science. Tel. +7 906 343-25-21. E-mail: o.zaharova@psuti.ru

Bednyak Svetlana Gennadievna, Povolzhskiy State University of Telecommunications and Informatics, 77, Moscovskoye shosse, Samara, 443090, Russian Federation; Associated Professor of Information Systems and Technologies Department, PhD in Pedagogical Sciences. Tel. +7 903 308-29-88. E-mail: s.bednyak@psuti.ru

References

1. Zakharova O.I. Development of a system for analysis and processing of text data. *Problemy tekhniki i tekhnologij telekommunikacij (PTiTT-2023): materialy XXV Mezhdunarodnoj nauchno-tekhnicheskoj konferencii*. Kazan': KNITU-KAI, 2023, pp. 261–262. (In Russ.)
2. Kuleshov S.V., Zaitseva A.A., Levashkin S.P. Technologies and principles of unstructured distributed data processing in the context of modern media content providing. *Informatizaciya i svyaz'*, 2020, no. 5. pp. 22–28. DOI: 10.34219/2078-8320-2020-11-5-22-28 (In Russ.)
3. Zakharova O.I. Semantic analysis and synthesis of text data. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyj analiz i informacionnye tekhnologii*, 2023, no. 4, pp. 182–208. DOI: 10.17308/sait/1995-5499/2023/4/182-208 (In Russ.)
4. Smetanin S., Komarov M. Deep transfer learning baselines for sentiment analysis in Russian. *Information Processing & Management*, 2021, vol. 58, no. 3, pp. 102484. DOI:10.1016/j.ipm.2020.102484. URL: https://www.sci-hub.ru/10.1016/j.ipm.2020.102484 (accessed: 28.06.2024).
5. Shavrina T.O. Methods of computational linguistics in the evaluation of artificial intelligence systems. *Voprosy yazykoznaniya*, 2021, no. 6, pp. 117–138. DOI: 10.31857/0373-658X.2021.6.117-138 (In Russ.)
6. Leshchinskaya N.M., Kolesnik M.A. Implementation of artificial intelligence technologies in Russia. *Sociologiya iskusstvennogo intellekta*, 2023, vol. 4, no. 2, pp. 63–72. (In Russ.)
7. Hartmann J. et al. Comparing automated text classification methods. *International Journals of Research Marketing*, 2019, vol. 36, no. 1, pp. 20–36. DOI:10.1016/j.ijresmar.2018.09.009. URL: https://www.sci-hub.ru/10.1016/j.ijresmar.2018.09.009 (accessed: 20.07.2024).
8. Liu Y. et al. A robustly optimized BERT pretraining approach. URL: https://arxiv.org/pdf/1907.11692.pdf (accessed: 28.07.2024).
9. Zakharova O.I., Levashkin S.P., Ivanov K.N. Modern Python libraries for collecting data from the Internet. *Problemy tekhniki i tekhnologij telekommunikacij (PTiTT-2020): materialy XXII Mezhdunarodnoj nauchno-tekhnicheskoj konferencii*, Samara: PSUTI, 2020, pp. 316–317. (In Russ.)
10. Ansari G.J. et al. A novel machine learning approach for scene text extraction. *Future Generation Computer Systems*, 2018, vol. 87, pp. 328–340. DOI: 10.1016/J.FUTURE.2018.04.074
11. Kim H. et al. Towards perfect text classification with Wikipedia-based semantic Naïve Bayes learning. *Neurocomputing*, 2018, vol. 315, pp. 128–134. DOI: 10.1016/J.NEUCOM.2018.07.002
12. Kratzwald B. et al. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 2018, vol. 115, pp. 24–35. DOI: 10.1016/J.DSS.2018.09.002
13. Taylor E.M. et al. Web opinion mining and sentimental analysis. *Advanced Techniques in Web Intelligence-2*, 2013, pp. 105–126. DOI: 10.1007/978-3-642- 33326-2_5
14. Wang Z. et al. A hybrid model of sentimental entity recognition on mobile social media. *EURASIP Journal on Wireless Communications and Networking*. DOI: 10.1186/s13638-016-0745-7. URL: https://sci-hub.ru/10.1186/s13638-016-0745-7 (accessed: 25.08.2024).
15. Altınel B., Ganiz M.C. Semantic text classification: A survey of past and recent advances. *Information Processing & Managemen*, 2018, vol. 54, no. 6, pp. 1129–1153. DOI: 10.1016/J.IPM.2018.08.001
16. Chatterjee A. et al. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 2019, vol. 93, pp. 309–317. DOI: 10.1016/J.CHB.2018.12.029

17. Mazharul S.M. et al. Sentiment analysis of tweet data. *Hoque Chowdhury*. URL: https://www.researchgate.net/publication/324965434_SENTIMENT_ANALYSIS_OF_TWEET_DATA (accessed: 27.4.2024).

18. Allahyari M. et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. URL: https://www.researchgate.net/publication/318336890_A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques (accessed: 30.08.2024).

19. Hedderich M.A. et al. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2545–2568.

20. Popovski G., Seljak B.K., Eftimov T. A survey of named-entity recognition methods for food information extraction. *IEEE Access*, 2020, vol. 8, pp. 31586–31594. DOI: 10.1109/ACCESS.2020.2973502

21. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. URL: https://arxiv.org/pdf/1810.04805 (accessed: 29.08.2024).

22. Canete J. et al. Spatish pre-trained BERT model and evaluation data. *Accepted as a workshop paper at PML4DC (ICLR)*. URL: https://www.researchgate.net/publication/372962444_Spanish_Pre-trained_BERT_Model_and_Evaluation_Data (accessed: 20.08.2024).

23. Ivanov K.N., Zakharova O.I. Natural language processing. application of language models. *Aktual'nye problemy informatiki, radiotekhniki i svyazi: materialy XXX Rossijskoj nauchno-tekhnicheskoj konferencii*. Samara: PSUTI, 2023, pp. 155–156. (In Russ.)

Received 05.08.2024

ТЕХНОЛОГИИ РАДИОСВЯЗИ, РАДИОВЕЩАНИЯ И ТЕЛЕВИДЕНИЯ

УДК 004.89

ИССЛЕДОВАНИЕ ВОЗМОЖНОСТЕЙ КОМПЬЮТЕРНОГО ЗРЕНИЯ В АВТОНОМНЫХ СИСТЕМАХ УПРАВЛЕНИЯ БЕСПИЛОТНЫХ ЛЕТАТЕЛЬНЫХ АППАРАТОВ

Карелин Е.А., Любащенко Т.Д., Палилов М.Р., Жиглова Н.С., Пачин А.В.
Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, Санкт-Петербург, РФ
E-mail: evgeniikarelin01@mail.ru, tima50879@gmail.com, palilovfox@gmail.com, zhiglova.natalia@yandex.ru, pachin.andrej@bk.ru

Развитие современных аппаратных и программных средств привело к стремительному распространению применения беспилотных аппаратов, в первую очередь летательных. Одним из перспективных направлений повышения эффективности подобных платформ является разработка для них систем автономного управления, исключающих участие человека-оператора. В статье представлены результаты исследования, целью которого является изучение возможности построения элементов автоматической системы управления движением на основе компьютерного зрения для беспилотных летательных аппаратов. Авторами использована методология сравнительного анализа для обоснования преимуществ одного из наиболее популярных инструментов – YOLO и SSD. Также выполнен сбор данных для обучения выбранной модели и ее тестирование в разных условиях. Описываются последовательность создания обучающего набора для эффективного обучения модели, а также результаты тестирования модели на видеоизображениях, полученных с камеры беспилотного летательного аппарата в различных условиях. Результаты тестирования подтверждают, что модель YOLOv8n пригодна для обнаружения объектов на борту беспилотного летательного аппарата с аппаратной платформой в виде одноплатного компьютера Raspberry Pi 4 Model B. Точность обнаружения объектов составила 80-90% при энергопотреблениии 15-25 Ватт.

Ключевые слова: *беспилотный летательный аппарат, компьютерное зрение, YOLO, машинное обучение, сверточная сеть, нейронная сеть, коптер, дрон*