6. ГОСТ Р 54878-2011. Аэрозоли преимущественно фиброгенного действия. URL: https://meganorm.ru/list2/64511-6.htm (дата обращения: 30.11.2023).

Об авторах:

МАКСИМОВ Руслан Андреевич – студент, ФГБОУ ВО «Тверской государственный технический университет», Тверь. E-mail: alcanara1234@yandex.ru

ТИХОНОВ Борис Борисович – кандидат технических наук, доцент кафедры биотехнологии, химии и стандартизации, ФГБОУ ВО «Тверской государственный технический университет» ТвГТУ, Тверь. E-mail: tiboris@yandex.ru

About the authors:

MAKSIMOV Ruslan Andreevich – Student, Tver State Technical University, Tver. E-mail: alcanara1234@yandex.ru

TIKHONOV Boris Borisovich – Candidate of Technical Sciences, Associate Professor of the Department of Biotechnology, Chemistry and Standardization, Tver State Technical University, Tver. E-mail: tiboris@yandex.ru

УДК 004.912

ПРОВЕДЕНИЕ СРАВНИТЕЛЬНОГО АНАЛИЗА МЕТОДОВ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДАННЫХ ПУТЕМ ИСПОЛЬЗОВАНИЯ РАЗРАБОТАННОЙ СИСТЕМЫ ПОИСКА НАУЧНОЙ ИНФОРМАЦИИ В СЕТИ ИНТЕРНЕТ

П.М. Трофимова

Аннотация. *В статье описан процесс тестирования разработанной на кафедре системы автоматизации поиска научной литературы определенной тематики в сети Интернет. Проведена оценка работы программного комплекса.*

Ключевые слова: *интеллектуальный анализ, веб-майнинг, веб-контент, точность, полнота, ошибка, аккуратность.*

Для тестирования модуля, разработанного применительно к программам Web Mining и TextStageProcessor, была выбрана тематика «центральный процессор персонального компьютера» на основе набора входных данных.

При тестировании программного комплекса по указанной тематике взяли следующие ключевые слова: «винчестер», «емкость», «накопитель», «время», «устройство», «модель», «компания», «скорость», enterprise, capacity, hdd.

Был сформирован список сайтов по указанной тематике. Ниже представлены соответствующие ссылки:

https://www.reg.ru/blog/chto-takoe-protsessor-cpu/
http://www.mediagnosis.ru/Autorun/Page6/5_3_.htm
https://works.doklad.ru/view/i8C8pD0Pkmk.html
https://dic.academic.ru/dic.nsf/ruwiki/1187537
https://works.doklad.ru/view/ufz4qakVXZ0.html
https://biosgid.ru/osnovy-ustrojstva-pk/processor-cpu-serdce-kompyutera.html
https://www.compuhome.ru/processor.html
https://prokompter.ru/centralnyj-processor-personalnogo-kompjutera/
https://ktonanovenkogo.ru/voprosy-i-otvety/processor-chto-ehto-takoe.html
https://digital-boom.ru/hardware/kak-vybrat-tsentralnyj-protsessor-kompyutera.html
https://club.dns-shop.ru/blog/t-100-protsessoryi/18597-kak-vyibrat-tsentralnyii-protsessor/
http://book.kbsu.ru/theory/chapter2/1_2_7.html
https://ezpc.ru/cpu1.shtml
https://htfi.ru/zhelezo/chto_takoe_centralnyj_processor.html
http://inep.sfedu.ru/wp-content/uploads/ehamt/learn/it/lection_5-6.pdf
https://mediapure.ru/matchast/chto-takoe-centralnyj-processor/
http://ravanda.ru/tests/3839
https://studme.org/62391/menedzhment/tsentralnyy_protsessor_mikroprotsessor
https://vuzlit.com/1032472/printsip_raboty_tsentralnogo_protsessora

Найденные поисковиком ссылки являются входными данными для программной системы поиска научной литературы. По этим входным данным система, разработанная авторами настоящей статьи, определяет сайты и анализирует их.

Набор обучающих данных для классификации создан на основе литературы, в частности научных статей по выбранной для тестирования теме. Данная выборка является обучающей для алгоритмов классификации текста.

С каждого сайта посредством веб-майнинга система извлекает текстовые данные и формирует на их основе файлы для последующего анализа методами text mining (рис. 1).

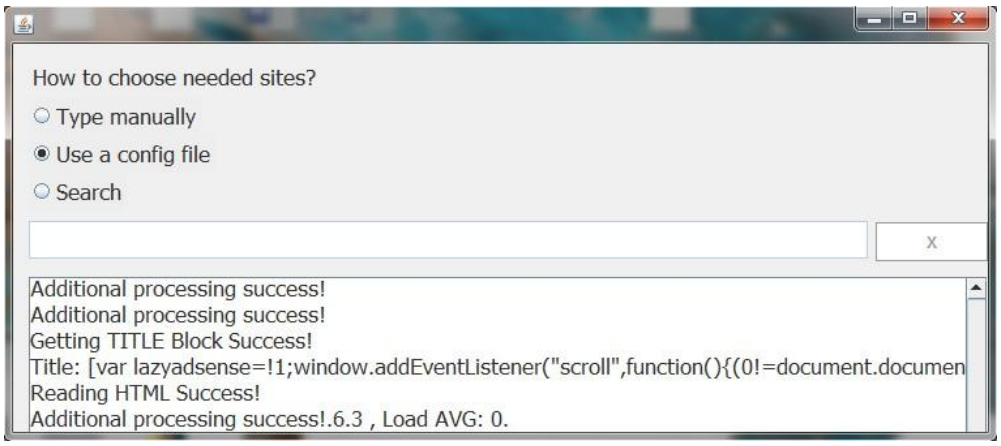

Рис. 1. Процесс извлечения текстовых данных с указанных сайтов

При ручной классификации ссылок можно установить, что по 3 ссылкам из 20 нет статей на выбранную тему. Получившийся результат будем считать первичной выборкой.

Далее в программе необходимо запустить алгоритм классификации на основе обучающей выборки и оценить результаты работы проекта (рис. 2).

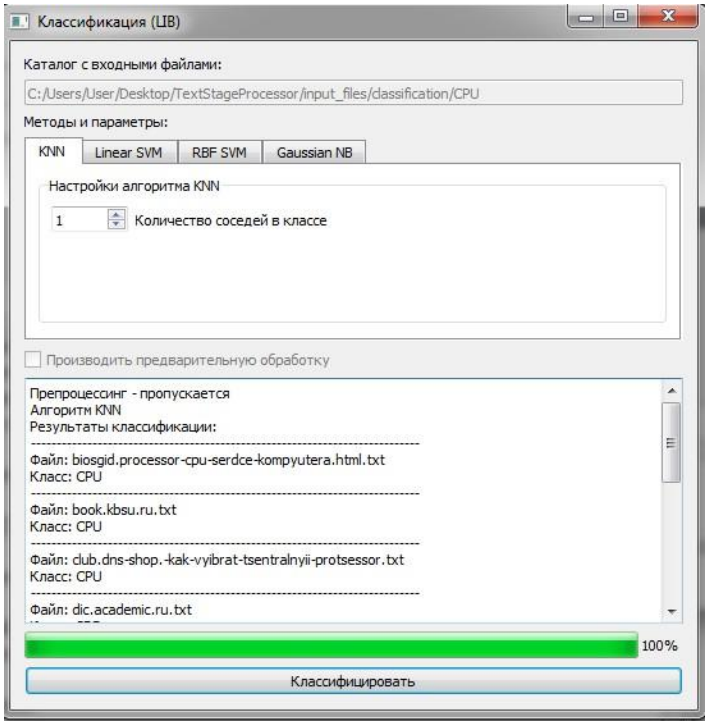

Рис. 2. Результаты классификации

Исходы классификации (таблица):

TP – true positive. Классификатор верно отнес объект к рассматриваемому классу.

TN – true negative. Классификатор верно указывает, что объект не принадлежит к рассматриваемому классу.

FP – false positive. Классификатор неверно отнес объект к рассматриваемому классу.

FN – false negative. Классификатор неверно указывает, что объект не принадлежит к рассматриваемому классу.

Оценка классификатора:

P **–** точность, т.е. доля истинно принадлежащих данному классу документов из всех, что классификатор записал в этот класс:

$$Prec = \frac{TP}{TP+FP}.$$

R **–** полнота, т.е. доля истинно принадлежащих данному классу документов и записанных в него классификатором среди всех документов, которые действительно ему принадлежат:

$$Recall = \frac{TP}{TP+FN}.$$

E – ошибка классификатора:

$$FPR = \frac{FP}{FP+TN}.$$

A **–** правильность (аккуратность) классификатора:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN}.$$

Confusion matrix (матрица несоответствий) для выборки из 20 ссылок

Показатель	Принадлежит классу (P)	Не принадлежит классу (N)
Предсказана принадлежность классу	13TP	0FP
Предсказано отсутствие принадлежности к классу	4FN	3TN

Точность = 13/(13 + 0) = 1.

Полнота = 13/(13 + 4) = 0,76.

Ошибка = 0/(0 + 3) = 0.

Правильность = (13 + 3)/(13 + 3 + 0 + 4) = 0,8.

Суммарная величина точности, полноты и правильности (I = P + R + A) = 2,56.

Таким образом, выборка на основе данных, полученных по результатам запуска программного кода, показала, что информация 13 из 20 сайтов соотносится с тематикой «центральный процессор персонального компьютера». При этом классификатор отнес объект к рассматриваемому

классу с точностью, равной единице. Это показывает, что разработанная программа упрощает анализ данных из сети Интернет, поскольку в выборке, полученной после ее запуска, все ссылки относятся к выбранной тематике.

Библиографический список

1. Logical Methods in Computer Science. URL: www.lmcs-online.org (дата обращения: 12.12.2023).
2. Психофизиология: искусственные нейронные сети. URL: scorcher/neuro/science/neurocomp/mem52. (дата обращения: 12.12.2023).
3. Анналин Ын, Кеннет Су. Теоретический минимум по Big Data. Все что нужно знать о больших данных // Библиотека программиста. СПб.: Питер, 2019. 208 с.
4. Беленький А. Текстомайнинг. Извлечение информации из неструктурированных текстов // КомпьютерПресс. 2008. № 10. URL: http://www.compress.ru/article.aspx?id=19605&iid=905 – 18.09.2011 (дата обращения: 11.12.2023).
5. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Е.И. Большакова [и др.]. М.: МИЭМ, 2011. 272 с.
6. Вандер Плас Дж. Python для сложных задач. Наука о данных и машинное обучение. СПб.: Питер, 2018. 576 с.
7. Панченко Е.Ю. Метод k средних при решении задачи распознавания диктора по речевому образцу // Молодой ученый. 2013. № 3 (50). С. 145–146.
8. Силен Д., Мейсман А. Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017. 336 с.

CONDUCTING A COMPARATIVE ANALYSIS OF CLUSTERING METHODS AND CLASSIFICATION OF TEXT DATA BY USING THE DEVELOPED SYSTEM FOR SEARCHING SCIENTIFIC INFORMATION ON THE INTERNET

P.M. Trofimova

Abstract. *The article describes the process of testing of the system developed at the department of automation of the search for scientific literature of a certain subject in the Internet. The work of the software system is evaluated.*

Keywords: *mining, web mining, web content, accuracy, completeness, error, accuracy.*

Об авторе:
ТРОФИМОВА Полина Михайловна – аспирант кафедры информационных технологий, ФГБОУ ВО «Тверской государственный технический университет», Тверь. E-mail: t.polina.m@gmail.com
Научный руководитель:
КАЛАБИН Александр Леонидович – доктор физико-математических наук, профессор, заведующий кафедрой программного обеспечения, ФГБОУ ВО «Тверской государственный технический университет», Тверь. E-mail: akalabin@yandex.ru

About the author:
Trofimova Polina Mikhailovna – Postgraduate Student of the Department of Information Technology, Tver State Technical University, Tver. E-mail: t.polina.m@gmail.com
Scientific adviser:
KHALABIN Aleksandr Leonidovich – Doctor of Physical and Mathematical Sciences, Professor, Head of the Department of Software, Tver State Technical University, Tver. E-mail: akalabin@yandex.ru

УДК 004.048

ЭКСПЕРТНО-АНАЛИТИЧЕСКИЕ ПОИСКОВЫЕ СИСТЕМЫ ДЛЯ БИЗНЕС-АНАЛИТИКИ. ЧАСТЬ 1: ДОСТОИНСТВА И НЕДОСТАТКИ

Т.Б. Яконовская, Л.В. Куликова,
В.Д. Славянский, Э.А. Арушанян

Аннотация. *В статье проведен обзор экспертно-аналитических поисковых систем, используемых в целях бизнес-разведки. Они позволяют аналитикам и специалистам по обработке данных извлекать значимую информацию из обширных наборов данных, улучшать маркетинговые исследования, поддерживать процесс принятия решений и эффективно управлять рисками. Однако, наряду с преимуществами, необходимо учитывать такие важные аспекты, как конфиденциальность данных и предвзятость алгоритмов. Отмечено, что указанные механизмы продолжают развиваться, а поэтому становятся значимыми инструментами для организаций, стремящихся преуспеть в бизнес-среде, основанной на данных.*