Торайғыров университетінің хабаршысы
ҒЫЛЫМИ ЖУРНАЛЫ

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Торайғыров университетінің ХАБАРШЫСЫ

Энергетикалық сериясы

1997 жылдан бастап шығады

ВЕСТНИК Торайгыров университета

Энергетическая серия

Издается с 1997 года

ISSN 2710-3420

№ 1 (2023)

ПАВЛОДАР

НАУЧНЫЙ ЖУРНАЛ
Вестник Торайгыров университета

Энергетическая серия
выходит 4 раза в год

СВИДЕТЕЛЬСТВО
о постановке на переучет периодического печатного издания,
информационного агентства и сетевого издания
№ 14310-Ж
выдано
Министерство информации и общественного развития
Республики Казахстан
Тематическая направленность
публикация материалов в области электроэнергетики,
электротехнологии, автоматизации, автоматизированных и
информационных систем, электромеханики и теплоэнергетики

Подписной индекс – 76136

https://10.48081/BNAS6555

SRSTI 06.73.65.

doi.org/10.48081/UPDW5717

A. S. Baimakhanova[1], K. M. Berkimbayev[2],
Eşref Adalı3, G. S. Iskendirova[4]
[1,2,4]Khoja Akhmet Yassawi International Kazakh-Turkish University,
Republic of Kazakhstan, Turkistan;
[3]Istanbul Technical University, Republic of Turkey, Istanbul
e-mail: aygerim.baymakhanova@ayu.edu.kz

IMPLEMENTING DOCUMENT CLASSIFICATION IN PYTHON AND EVALUATING RESULTS

In our article, we will cover general information about machine learning, its main types, as well as the most important libraries for machine learning in Python. Machine learning is one of the main methods for demonstrating data science in general. In machine learning, the computational and algorithmic capabilities of data science are combined with approaches, and the result is a set of data mining approaches, mainly related to the efficiency and theory of computation. Document classification plays an important role in the archiving department, they classify pre-defined documents and store them in a digital archive. Relevance of the topic In digital archives, document classification is an important process for many document preservation organizations. In the course of the study, a technology for testing the use of deep learning algorithms is being created. Deep learning, i.e. deep structured learning, is part of a broad group of machine learning methods based on artificial neural networks. Learning cannot be controlled, partially controlled or controlled.

Keywords: Machine learning, Artificial neural networks, Python, Scikit-learn, Keras, TensorFlow.

Introduction

«Machine learning» this term sometimes effectively solves all data problems. Although the possibilities of these methods are enormous, they must be used effectively. Machine learning is often considered part of the field of artificial intelligence. The program is trained on real data and can then be used to predict and understand various aspects of the data from new observations. At a basic level,

machine learning can be viewed in two main ways. Machine learning with the help of a teacher called be Supervised learning, without the help of a teacher machine learning called be Unsupervised learning. Machine learning for teachers involves modeling data labels and corresponding labels. Once a model is selected, it can be used to label new, previously unknown data. It is divided into classification and regression tasks. In classification, labels are discrete categories, while in regression they are continuous variables. Untrained machine learning involves modeling the features of a dataset without any features. These models include tasks such as clustering and dimensionality reduction. Clustering algorithms are used to extract individual groups of data, while reduction algorithms are designed to find compressed representations of data. There are partial teaching methods that combine the advantages and disadvantages of the two original methods.

Deep learning, i.e. deep structured learning, is part of a broad group of machine learning methods based on artificial neural networks. Learning cannot be supervised, partially supervised or supervised [1].

Document classification is an important task in the office, which is designed to classify pre-defined documents and store them in a digital archive. Document classification is an important process for many organizations when storing digital documents. In the course of the study, a technology for testing the use of Deep learning algorithms is being created.

If we consider the stages of classification on three grounds, according to the location of documents, analysis of textual information, content [2].

When classifying our documents, textual and visual are divided. NLP is used in sentiment analysis to analyze words and phrases in research [3].

Currently, deep learning convolutional neural networks are becoming a powerful tool for recognizing CNN layers.

This research focuses on the use of deep learning for document classification. In this study, a complex classification and compilation of documents is performed. In CNN deep learning, eight pre-trained layers prepared by ImageNet are trained from scratch on a dataset and then trained and classified to train the CNN.

This study examined the effect of image types on the accuracy of deep learning and transfer learning, as well as the effect of image types on crack level classification. The results show that the CNN models have the highest accuracy on the pooled image during deep zero training. In deep learning methods, Convolutional Neural Networks are known to use CNNs for computer tasks. The role of CNN in image classification is huge, it is modern and competitive, effective for complex work [4].

Unlike video analysis methods, CNN does not require manual generation of rules and automatically extracts multilevel features [5].

Materials and methods

Python is an object-oriented language with multiple inheritance. We can say that Python supports the classic OO model with some peculiarities. Classes in Python can have static variables that are common to all instances of the class, but cannot have static methods. All methods are class instances.

The Python programming language appeared in 1991 and quickly gained popularity and became a recognized language for machine learning. It is characterized by the ease of use of high-level programming languages. Learn how to load and visualize data, perform statistical calculations, process images, and more in Python. There are libraries for Let's take a closer look at the most popular Python machine learning libraries: Scikit-learn, PyTorch, Caffe, TensorFlow, Keras, OpenCV and TensorFlow [6].

Scikit-learn is one of the oldest and once popular libraries for training neural networks. It was created in 2007 and is still a reliable tool in areas such as classification, regression, clustering, modeling.

Keras is an open source library written in Python for interacting with artificial neural networks.

Basic principles of the scikit-learn library's statistical evaluation API:

- uniformity lies in the fact that all objects have the same interface and are based on a limited set of methods;

- visibility of the given values of control parameters as public attributes;

- limited object hierarchy, Python language classes are used only for algorithms, data sets are presented in standard formats;

- the integration of many machine learning problems can be expressed as a sequence of low-level algorithms;

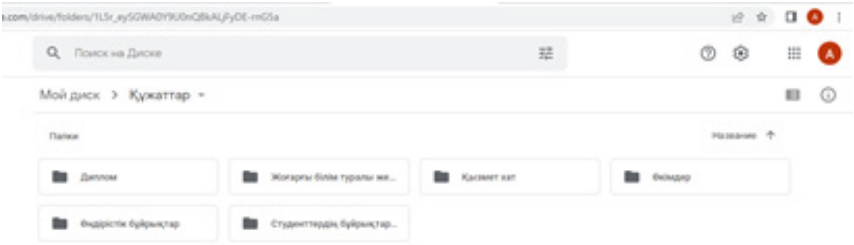
- sensible values sets appropriate initial values for option templates required by the library. In practice, these principles make learning the Scikit-learn library easier.

Often, using the Scikit-learn Statistical Estimation API involves the following steps: a model class is selected by importing the appropriate class from the Scikit-learn library; select the hyperparameters of the model by creating an instance of this class with the appropriate values; put data into feature matrix and target vector; train the model on your data; applying the model to new data.

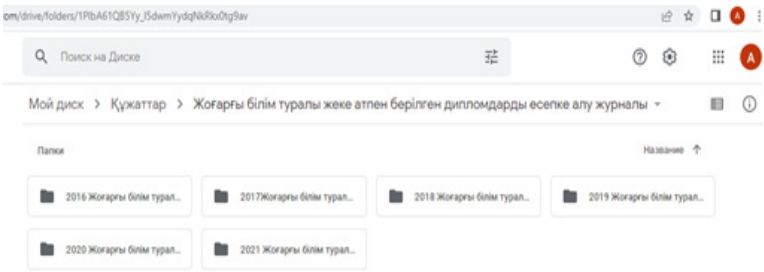
TensorFlow - is currently the most popular machine learning library, meaning it is a very useful library for our work. It has a simple intuitive interface that facilitates the implementation of neural networks, is ideal for developing complex projects such as building multi-layer neural networks, and its training methods are constantly being improved. Let's take a closer look at the architecture of the TensorFlow library. It works with a static calculation schedule. First, the graph is set, then the calculations begin, if necessary, changes are made to the architecture and the model is retrained. This approach was chosen for reasons of efficiency, but

many modern machine learning tools are able to take into account changes in the learning process without a significant loss in speed. The TensorFlow library can be used to perform numerical calculations. In this library, these calculations are done using graphs called data flow. In these graphs, the vertices are mathematical operations, and the edges are data, usually represented as multidimensional arrays or tensors connected by these edges. The name «TensorFlow» comes from artificial neural network computations with multidimensional data and tensors, literally «tensor flow». Tensors are the core objects in TensorFlow and are implemented as n-dimensional arrays of data that allow you to represent data in complex dimensions. Each dimension can be thought of as a separate label [7].

Our documents are saved to disk. In our study, 6 classes were identified and supplemented.

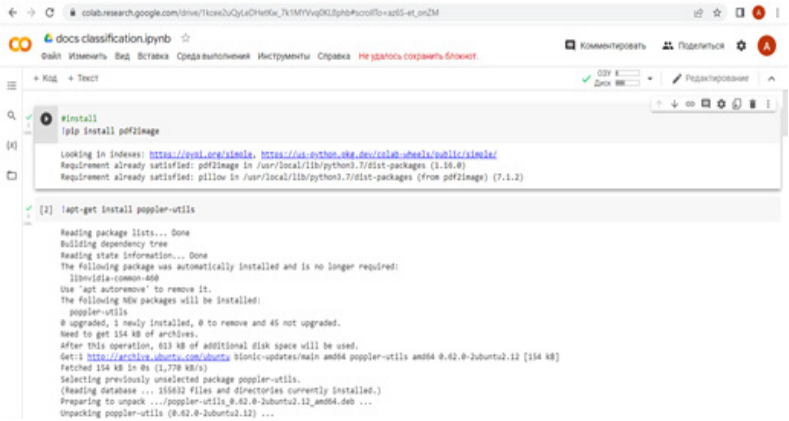


Picture 1 –The number of classes of our work is planned, at the moment 6 classes have been recruited.



Picture 2 – Registration of diplomas of higher education issued in an individual name

The existing algorithms and strategies of the classification task allow to reduce the error in half of the document image [8].Let's take a look at our exploration of these collected documents using the Python programming language. Let's write a function that calculates the confidence level of the scanned text, further illustrating our work in the Python programming language:AlexNet is a convolutional neural network that has greatly influenced the development of computer vision machine learning algorithms. In 2012, The Big Network won the IMAGENET lsvrc-2012 image recognition competition (15.3 % error rate, 26.2 % runner-up). The architecture of AlexNet is similar to the LeNet network created by Yann LeCum. However, AlexNet has many per-layer filters and built-in convolutional layers. The network includes summation, maximum pooling, collapse, data pooling, ReLU activation functions, and stochastic gradient descent.



Picture 3 –I mage recognition

Features of AlexNet. Relu is used instead of the arc tangent as an activation to add non-linearities to the model. Due to this, the speed is 6 times higher than the accuracy of the method.

Results and discussion

Using a drip instead of a corrector solves the problem of retraining. However, the reading time is doubled with a drop rate of 0.5.

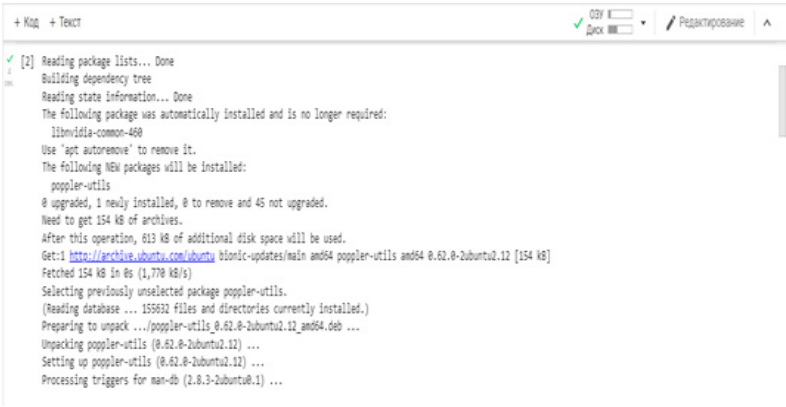
Unions are closed to reduce the size of the network. Due to this, the error rate at the first and fifth levels will decrease to 0.4 percent and 0.3 percent, respectively.

A convolutional neural network (CNN) differs from the other two models in that each layer in a CNN has a known convolution operation, hence the name «convolutional» neural network. The weights in CNNs are distributed among neurons, similar to the communication patterns between neurons in the animal

visual cortex, which was apparently a network inspired by a biological process. Typically, the results of multiple convolutional layers used sequentially in a CNN are reduced with a pooling layer to speed up the process. Thus, a maximum pooling layer or a global pooling layer is often added after the convolutional layers.

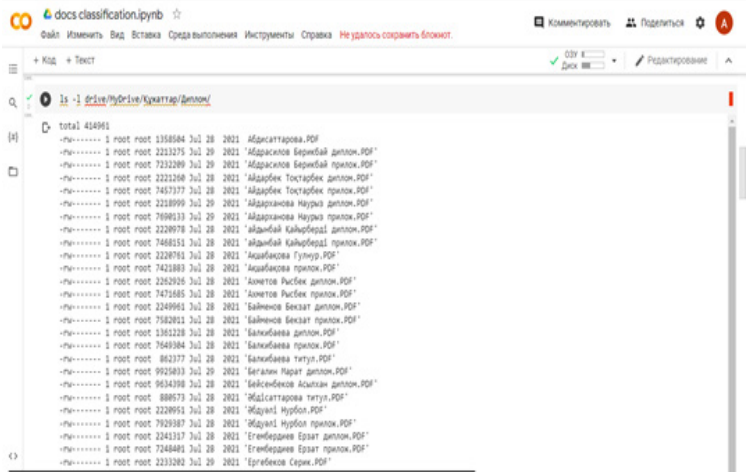
The results obtained using various deep learning methods and simple machine learning methods are compared. It also includes a comparative analysis of the effectiveness of various submission methods.

Switch to the main function: image scanning:



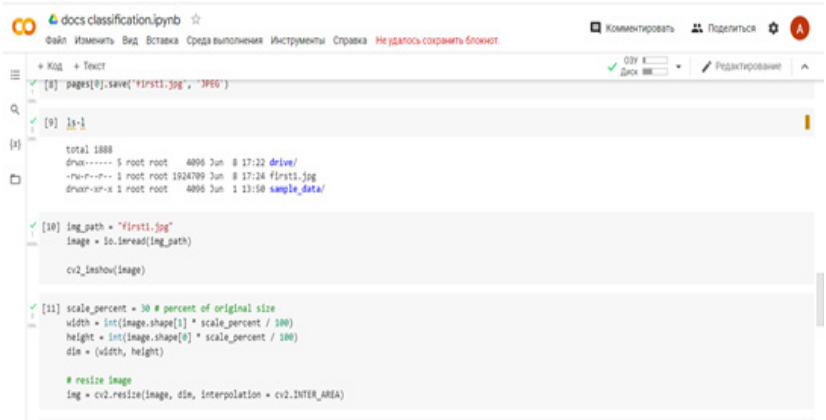
Picture 4 –The process of checking the documents we have collected

Picture 5– Lists of students are shown.



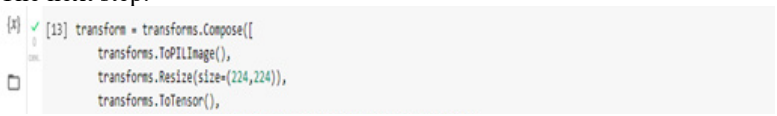
Picture 6 – List

It was then extended to the complex task of generating features (signatures) for image classification, often constructed as a combination of CNN and LSTM [9].



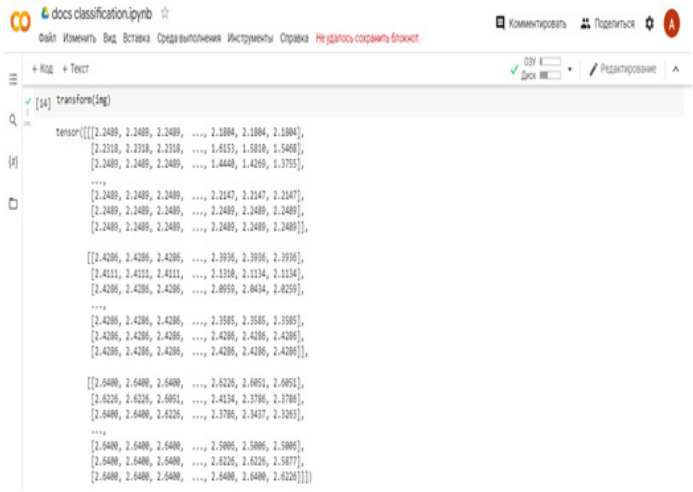
Picture 7–The result of our created program

The next step.



Picture 8–Stage of digitalization

Unlike video analysis methods, CNN does not require manual classification of documents and can automatically extract the required features of the layers [10].



Picture 9 – Stage of digitalization

The above will do the following:

– Scans an image buffer or image file.

– Pre-processes the image.

– Starts the Tesseract engine with predefined options.

- Calculates the confidence level of the received image content.

– Draws a green bar around readable text elements with a confidence score greater than 30.

– Search for specific text in captured image content.

– Highlights or edits found matches with the search text.

– Readable text fields, displays a window with selected or edited text.

– Creates the text content of an image.

– Prints a summary to the console.

So, one of the most popular tools for machine learning is Python. In our study, the Python programming language allows us to work with many machine learning libraries such as Scikit-learn, PyTorch, Caffe, TensorFlow, OpenCV.

Conclusions

It must be said that machine learning has become an integral part of our lives, from medical diagnosis to subsequent treatment and social networking. One of the most popular machine learning tools is Python, which combines power and

ease of use. It allows you to work with many machine learning libraries such as Scikit-learn, PyTorch, Caffe, TensorFlow, OpenCV.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 **Alladi, T., Kohli, V., Chamola, V., Yu, F.R., Fellow,** IEEE. A deep learning based misbehavior classification scheme for intrusion detection in cooperative intelligent transportation systems [Text] // Digital Communications and Networks. –2022. – P. 2–15.

2 **Hu, J., Kashi, R.,** Wilfong, G. Document image layout comparison and classification [Text]//Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR'99 (Cat. No. PR00318). – IEEE, 1999. – P. 285–288.

3 **Kang, L., Kumar, Dzh., Je, P., Li, YU., Doermann, D.** Convolutional neural networks for document image classification [Text] //2014 22nd International Conference on Pattern Recognition. – IEEE, 2014. – P. 3168–3172.

4 **LeCun, Y., Bengio, Y., Hinton, G.** Deep learning [Text] //nature. – 2015. – T. 521. – №. 7553. – P. 436–444.

5 **Lyu, Z., Cao, YU., Van, YU., Van, V.** Computer vision-based concrete crack detection using U-net fully convolutional networks [Text] //Automation in Construction. – 2019. – T. 104. – P. 129–139.

6 **Hu, B., Ergu, D., Yan, H., Lyu, K., Kaj, YU.** Document images classification based on deep learning [Text] //Procedia Computer Science. – 2019. – T. 162. – P. 514-522.

7 **Gupta, M., Bedi, P., Dzhagvani, P., Bhasin, V.** Gene Mutation Classification through Text Evidence Facilitating Cancer Tumour Detection [Text] //Journal of Healthcare Engineering. – 2021. – T 2021.

8 **Afzal, M. Z., Kyol'sh, A., Ahmed, S., Livicki, M.** Cutting the error by half: Investigation of very deep cnn and advanced training strategies for document image classification [Text] //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – IEEE, 2017. – T. 1. – P. 883–888.

9 **Cireşan, D., Dzhusti, A., Gambardella, L. M., SHmidhuber, Dzh.** Mitosis detection in breast cancer histology images with deep neural networks [Text] // International conference on medical image computing and computer-assisted intervention. – Springer, Berlin, Heidelberg, 2013. – P. 411–418.

10 **Schmidhuber, J.** Deep learning in neural networks: An overview [Text] //Neural networks. – 2015. – T. 61. – P. 85–117.

REFERENCES

1 **Alladi, T., Kohli, V., Chamola, V., YU, F.R.**, nauchnyj sotrudnik IEEE. Skhema klassifikacii narushenij povedeniya na osnove glubokogo obucheniya dlya obnaruzheniya vtorzhenij v kooperativnye intellektual'nye transportnye sistemy [A deep learning based misbehavior classification scheme for intrusion detection in cooperative intelligent transportationsystems] // Cifrovye kommunikacii i seti [Digital Communications and Networks]. – 2022. – R. 2–15. [in English]

2 **Hu, J., Kashi, R., Wilfong, G.** Sravnenie i klassifikaciya komponovki izobrazheniya dokumenta [Document image layout comparison and classification] // V materialah Pyatoj mezhdunarodnoj konferencii po analizu i raspoznavaniyu dokumentov [In Proceedings of the Fifth International Conference on Document Analysis and Recognition]. – Indiya, Bangalor, 1999, sentyabr'. – P. 285–288. [in English]

3 **Kang, L., Kumar, Dzh., Je, P., Li, YU., Doermann, D.** Svertochnye nejronnye seti dlya klassifikacii izobrazhenij dokumentov [Convolutional neural networks for document image classification] // V 2014 g. 22-ya Mezhdunarodnaya konferenciya po raspoznavaniyu obrazov [In 2014 22nd International Conference on Pattern Recognition]. – SHveciya, Stokgol'm, 2014, avgust. – P. 3168–3172. [in English]

4 **LeCun, Y., Bengio, Y., Hinton, G.** Glubokoe obuchenie [Deep Learning] // Priroda [Nature]. – 2015. – № 521. – P. 436–444. [in English]

5 **Lyu, Z., Cao, YU., Van, YU., Van, V.** Obnaruzhenie treshchin v betone na osnove komp'yuternogo zreniya s ispol'zovaniem polnost'yu svertochnyh setej U–net [Computer vision-based concrete crack detection using U–net fully convolutional networks] // Avtomatizaciya v stroitel'stve [Automation in Construction]. – 2019. – T. 104. – P. 129–139 [in English].

6 **Hu, B., Ergu, D., Yan, H., Lyu, K., Kaj, YU.** Klassifikaciya izobrazhenij dokumentov na osnove glubokogo obucheniya [Document images classification based on deep learning] // Procedia Computer Science. – 2019. – T. 162. – P. 514–522 [in English]

7 **Gupta, M., Bedi, P., Dzhagvani, P., Bhasin, V.** Klassifikaciya gennyh mutacij s pomoshch'yu tekstovyh dokazatel'stv, oblegchayushchih obnaruzhenie rakovyh opuholej [Gene Mutation Classification through Text Evidence Facilitating Cancer Tumour Detection] // Healthcare Engineering [Healthcare Engineering]. –2021. – P. 233–250 [in English].

8 **Afzal, M. Z., Kyol'sh, A., Ahmed, S., Livicki, M.** Sokrashchenie oshibki vdvoe: issledovanie ochen' glubokoj CNN i strategii povysheniya kvalifikacii dlya klassifikacii izobrazhenij dokumentov [Cutting the Error by Half: Investigation of Very Deep CNN and Advanced Training Strategies for Document Image Classification] // 14-ya Mezhdunarodnaya konferenciya IAPR po analizu i raspoznavaniyu dokumentov [14th IAPR International Conference on Document

Analysis and Recognition]. – Germaniya, Universitet Kajzerslauterna, 2017. aprel'. – R. 1–6 [in English].

9 **Cireşan, D., Dzhusti, A., Gambardella, L. M., SHmidhuber, Dzh.** Obnaruzhenie mitoza na gistologicheskih izobrazheniyah raka molochnoj zhelezy s ispol'zovaniem glubokih nejronnyh setej [Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks] // Proceedings MICCAI. Konspekt lekcij po informatike [Proceedings MICCAI. Lecture Notes in Computer Science]. – 2013. – P. 411–418 [in English].

10 **Schmidhuber, J.** Glubokoe obuchenie v nejronnyh setyah: obzor [Deep Learning in Neural Networks: An Overview] // Nejronnye seti [Neural Networks]. – 2015. – P. 85–117 [in English].

Material received on 13.03.23

**А. С. Баймаханова[1], К. М. Беркимбаев[2], Е. Адалы[3], Г. С. Искендірова[4]*
[1,2,4]Қожа Ахмет Ясауи атындағы Халықаралық қазақ-түрік университеті,
Қазақстан Республикасы, Түркістан қ.;
[3]Стамбул техникалық университеті,
Түркия Республикасы, Стамбул қ.
Материал баспаға 13.03.23 түсті.

PYTHON БАҒДАРЛАМАСЫНДА ҚҰЖАТТАРДЫ КЛАССИФИКАЦИЯЛАУДЫҢ ЖҮЗЕГЕ АСЫРЫЛУЫ ЖӘНЕ НӘТИЖЕЛЕРДІҢ БАҒАЛАНУЫ

Біздің жұмысымызда машиналық оқыту туралы жалпы мәліметтер, оның негізгі түрлері, сондай-ақ Python тілінде машиналық оқытуға арналған ең маңызды кітапханалар қарастырылады. Машиналық оқыту – бұл жалпы деректер ғылымын көрсетудің негізгі әдістердің бірі болып табылады. Машиналық оқытуда деректер ғылымының есептеу және алгоритмдік мүмкіндіктерін тәсілдермен біріктіріліп, нәтижесі негізінен тиімділік пен есептеу теориясымен байланысты деректерді зерттеу тәсілдерінің жиынтығы. Құжаттарды жіктеу архив бөлімінде маңызды рөл атқарады, олар алдын-ала анықталған құжаттарды жіктеп және цифрлық мұрағатта сақтайды. Тақырыптың өзектілігі цифрлық мұрағаттарда көптеген құжаттарды сақтау ұйымдары үшін құжаттарды жіктеу маңызды процесс болып табылады. Зерттеу барысында Deep Learning алгоритмдерін қолдануды дамыту технологиясы жасалады. Терең оқыту, яғни терең құрылымдық оқыту – бұл жасанды нейрондық желілерге негізделген машиналық

оқыту әдістерінің кең тобының бөлігі болып табылады. Оқытуды бақылауға, ішінара бақылауға немесе бақылауға болмайды.

Кілтті сөздер: машиналық оқыту, жасанды нейрондық желілер, Python, Scikit-learn, Keras, TensorFlow.

**А. С. Баймаханова[1], К. М. Беркимбаев[2], Е. Адалы[3], Г. С. Искендирова[4]*
[1,2,4]Международный казахско-турецкий университета
имени Ходжи Ахмеда Ясави, Республика Казахстан, г. Туркестан;
[3]Стамбульский технический университет,
Турецкая Республика, г. Стамбул.
Material received on 13.03.23

ВНЕДРЕНИЕ КЛАССИФИКАЦИИ ДОКУМЕНТОВ В PYTHON И ОЦЕНКА РЕЗУЛЬТАТОВ

В нашей статье мы рассмотрим общую информацию о машинном обучении, его основных видах, а также самых важных библиотеках для машинного обучения в Python. Машинное обучение – один из основных методов демонстрации науки о данных в целом. В машинном обучении вычислительные и алгоритмические возможности науки о данных сочетаются с подходами, и в результате получается набор подходов к интеллектуальному анализу данных, в основном связанных с эффективностью и теорией вычислений. Классификация документов играет важную роль в архивном отделе, они классифицируют заранее определенные документы и хранят их в цифровом архиве. Актуальность темы В цифровых архивах классификация документов является важным процессом для многих организаций по сохранению документов. В ходе исследования создается технология тестирования использования алгоритмов глубокого обучения. Глубокое обучение, то есть глубокое структурированное обучение, является частью широкой группы методов машинного обучения, основанных на искусственных нейронных сетях. Обучение нельзя контролировать, частично контролировать или контролировать.

Ключевые слова: машинное обучение, искусственные нейронные сети, Python, Scikit-learn, Keras, TensorFlow.

Теруге 13.03.2023 ж. жіберілді. Басуға 31.03.2023 ж. қол қойылды.
Электронды баспа
3,44 Mb RAM
Шартты баспа табағы 23.59. Таралымы 300 дана. Бағасы келісім бойынша.
Компьютерде беттеген: А. К. Мыржикова
Корректор: А. Р. Омарова
Тапсырыс № 4039
Сдано в набор 13.03.2023 г. Подписано в печать 31.03.2023 г.
Электронное издание
3,44 Mb RAM
Усл. печ. л. 23.59. Тираж 300 экз. Цена договорная.
Компьютерная верстка: А. К. Мыржикова
Корректор: А. Р. Омарова, Д. А Кожас
Заказ № 4039

«Toraighyrov University» баспасынан басылып шығарылған
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.

«Toraighyrov University» баспасы
Торайғыров университеті
140008, Павлодар қ., Ломов к., 64, 137 каб.
67-36-69
E-mail: kereku@tou.edu.kz
www.vestnik-energy.tou.edu.kz