MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA

Universitatea Tehnică a Moldovei

Facultatea Calculatoare, Informatică și Microelectronică

Departamentul Ingineria Software și Automatica

Admis la susținere
Șef departament:
Ion Fiodorov, conf. univ., dr.

_____
“____” _____ 2024

Managementul datelor prin aplicarea algoritmilor de învățare automată

Proiect master

Student: _____ **Rotaru Vasile, TI-221M**

Coordonator: _____ **Duca Ludmila, asist. univ.**

Consultant: _____ **Cojocaru Svetlana, asist. univ.**

Chișinău 2024

ADNOTARE

Această teză explorează intersecția dintre date masive, data mining și învățarea automată prin prisma tehnologiilor bazate pe *Python*, în special *Pandas, scikit-learn* și *Langchain*. Studiul este structurat în trei capitole, fiecare abordând un aspect crucial al temei generale de cercetare.

Capitolul 1 este despre date masive și data Mining. Acest capitol oferă o prezentare cuprinzătoare a peisajului big data și data mining, evidențiind provocările și oportunitățile generate de creșterea exponențială a datelor în domenii diverse. Accentul este pus pe înțelegerea fundațiilor teoretice ale tehnicilor de data mining și a aplicațiilor practice în extragerea de informații valoroase din seturi de date mari. Capitolul explorează, de asemenea, importanța procesării eficiente și a managementului datelor în contextul analizei big data.

Capitolul 2 este despre *Python, Pandas* și *scikit-learn*, concentrându-se pe implementarea practică a analizei datelor, acest capitol se adâncește în capacitățile limbajului de programare *Python* și ale bibliotecilor sale larg utilizate - *Pandas* și *scikit-learn*. Se examinează modul în care aceste instrumente facilitează manipularea eficientă a datelor, explorarea și aplicarea algoritmilor de învățare automată pentru analiză predictivă.

În capitolul 3 se relatează despre structura sistemului proiectat pentru analiza datelor. Se iau cazurile pentru sistemul de returnare a informației. Acestea vizează și modele lingvistice largi, dar și date prestate de Biroul Național de Statistică al Moldovei. Datele sunt colectate cu ajutorul unui crawler web implementat specific pentru a parcurge datele din această sursă. Ulterior sunt descrise etapele de curățare și organizare a datelor. Mai mult, capitolul 3 expune procesul de utilizare a modelelor lingvistice largi pentru a analiza datele, dar și utilizarea unui model lingvistic creat de la zero.

Relatările capitolului 4 se referă la rezultatele obținute în urma proiectării și realizării sistemului informatic descrise în capitolul 3 pe baza datelor acumulate pe baza datelor acumulate corespunzător descrierii expuse la fel în capitolul 3. În cadrul capitolului sunt aduse exemple de performanță, precizie și cazuri de utilizare a sistemului. Mai mult, sunt făcute și comparații între modelele lingvistice largi care sunt utilizate ca parte dinamică a sistemului informatic.

Prin această explorare expusă în mai multe capitole, teza își propune să demonstreze sinergia dintre domeniile datelor masive, data mining și implementarea tehnică în limbajul *Python*, culminând prin integrarea inovatoare a *Langchain* pentru a avansa în domeniul analizei predictive. Concluziile prezentate aici contribuie la peisajul în continuă evoluție al științei datelor, oferind perspective practice pentru cercetători, practicieni și organizații care doresc să valorifice în întregime potențialul datelor utilizate.

ANNOTATION

This thesis explores the intersection of big data, data mining, and machine learning through the lens of *Python*-based technologies, specifically *Pandas*, *scikit-learn*, and *Langchain*. The study is structured into three chapters, each addressing a crucial aspect of the overall research theme.

Chapter 1 focuses on big data and data mining. This chapter provides a comprehensive overview of the big data and data mining landscape, highlighting the challenges and opportunities posed by the exponential growth of data in various fields. The emphasis is on understanding the theoretical foundations of data mining techniques and their practical applications in extracting valuable insights from large datasets. The chapter also explores the importance of efficient data processing and management in the context of big data analytics.

Chapter 2 delves into *Python*, *Pandas*, and *scikit-learn*, concentrating on the practical implementation of data analysis. This chapter explores the capabilities of the *Python* programming language and its widely used libraries—*Pandas* and *scikit-learn*. It examines how these tools facilitate efficient data manipulation, exploration, and the application of machine learning algorithms for predictive analytics.

In Chapter 3, the structure of the system designed for data analysis is discussed. The cases for the information retrieval system are considered, targeting both large language models and data provided by the National Bureau of Statistics of Moldova. The data is collected using a web crawler implemented specifically to navigate through this source. Subsequently, the steps of data cleaning and organization are described. Furthermore, chapter 3 outlines the process of using large language models to analyze the data, as well as the use of a language model created from scratch.

Chapter 4 reports the results obtained from the design and implementation of the computer system described in chapter 3, based on the data accumulated corresponding to the description outlined in chapter 3. The chapter provides examples of performance, accuracy, and use cases of the system. Moreover, a comprehensive comparison is made between the large languages models used as a dynamic part of the computer system.

Through this exploration across multiple chapters, the thesis aims to demonstrate the synergistic power of big data, data mining, and Python-based technologies, culminating in the innovative integration of Langchain to advance the field of predictive analytics. The conclusions presented here contribute to the ever-evolving landscape of data science, providing practical perspectives for researchers, practitioners, and organizations looking to fully leverage the potential of their data assets.

CUPRINS

INTRODUCERE

Momentan lumea trece printr-o ascensiune tehnologică foarte rapidă. O mare parte din tehnologiile noi sunt niște descoperiri vechi, dar îmbunătățite și lustruite pentru o mai bună experiență în utilizare. Totuși, nu este atât de ușor de a crea algoritmi inteligenți și suficient de flexibili pentru orice domeniu de aplicare. În așa caz se poate de spus că este nevoie de experiență pentru crearea unor astfel de algoritmi universali. Cea mai mare parte a experienței se acumulează din datele despre situațiile deja petrecute. Deci pentru a crea algoritmi flexibili este nevoie de date.

Într-adevăr, ascensiunea tehnologică din prezent se datorează datelor și a „experienței" acumulate din datele colectate. În ziua de azi sau dezvoltat intens și „senzorii" ce continuu colectează datele, acestea ajungând la ordinul exaocteților EB sau chiar zettaocteților ZB. În așa cazuri se aplică strategii din domeniul datelor masive, adică Big Data.

Datorită faptului că algoritmii proiectați până în prezent sunt suficient de flexibili pentru aplicarea în domeniile existente, ar fi bine aplicarea acestora în domenii mai specifice precum piața muncii sau educația dintr-o regiune specifică sau chiar dintr-o țară. O mulțime de strategii de prelucrare a datelor, dar și algoritmi de analiză a lor sunt des aplicate în domenii precum marketing, prognozarea pieței financiare, prognozarea evoluției meteo, etc. Într-o analogie, se poate de aplicat algoritmi similari și în domeniul pieței muncii, ca spre exemplu sisteme de recomandare, sau chiar baze de cunoștințe în tandem cu sistem de echilibrare. O metoda, la fel destul de ambițioasă, ar fi rețele neuronale. Pentru a generaliza, scopul ar fi de a avea un instrument ce va oferi posibilitatea de aplicarea a diferitor procedee de analiză în timp real cu utilizarea modelelor lingvistice largi (LLMs). Un mic exemplu ar fi construcția graficului regresiei lineare pentru un set de date specific.

Pe baza unui LLM specific proiectat pentru analiza de date se pot construi sisteme mai complexe care vor putea fi utilizate cu facilitate în mai multe studii de fezabilitate, dar și în domenii de analiză unde s-ar putea să fie nevoie de aplicarea procesului de analiză a datelor pe post de scurtă incursiune în detaliile oferite de datele colectate. Spre exemplu, se dorește o scurtă detaliere a evoluție ratei șomajului în Republica Moldova și sugestii de metode pentru prognozarea acesteia într-o perioadă imediat apropiată. Pentru acest exemplu elocvent se poate spune că e situația unei scurte detalieri a domeniului și careva sugestii pentru un posibil studiu de fezabilitate. Un alt exemplu mai complex ar fi impactul nivelului educației asupra ratei șomajului, unde este necesară analiza nu doar a datelor din educație, dar și a datelor din domeniul muncii. Ca etapă finală este de a corela cele două domenii prin aplicarea unui proces de inter-analiză, precum și o analiză a domeniul corelat din punctul de vedere a domeniului curent.

Teren de experimentare este suficient, totuși pentru un rezultat bun mai este necesar ca și „nucleul" datelor să fie suficient de calitativ și valoros. Astfel de sisteme ar putea să reprezinte un mod de agregare, dar și un mod de echilibrare a domeniului ales pentru datele prelucrate.

BIBLIOGRAFIE

[1] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Indianapolis: Wiley, 2015.

[2] „Smartphone Icons, Logos, Symbols - Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/5wGnhtHODuE9/mobile-phone.

[3] „Social network Icons, Logos, Symbols - Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/113702/social-network.

[4] „Video monitoring Icons, Logos, Symbols – Free Download PNG, SVG," 26 11 2023. [Interactiv]. Available: https://icons8.com/icon/16200/bullet-camera.

[5] „Video supervision Icons, Logos, Symbols – Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/V6jsd74jR29Q/video.

[6] „Intelligent mapping Icons, Logos, Symbols – Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/2kHyhcs10JJa/intelligent-website.

[7] „Radio tower Icons, Logos, Symbols – Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/FK6THEZ6T2FW/radio-tower.

[8] „Brain Icons, Logos, Symbols – Free Download PNG, SVG," [Interactiv]. Available: https://icons8.com/icon/tXnJ4Th5bdzX/brain.

[9] „Dna Icons, Logos, Symbols – Free Download PNG, SVG," [Interactiv].

[10] C. C. Aggarwal, Data Mining: The Textbook, New York: Springer, 2015.

[11] S. Mukhopadhyay, Advanced Data Analytics Using Python, Kolkata, West Bengal, India: Apress, 2018.

[12] M. R. Douglas, „Large Language Models," nr. 2307.05782v2, 2023.

[13] „The Python Tutorial," [Interactiv]. Available: https://docs.python.org/3/tutorial/index.html.

[14] „Pandas Library documentation," [Interactiv]. Available: https://pandas.pydata.org/docs/.

[15] „Using Matplotlib — Matplotlib 3.8.2 documentation," [Interactiv]. Available:

https://matplotlib.org/stable/users/index.

[16] „Introduction | Langchain,” [Interactiv]. Available: https://python.langchain.com/assets/images/langchain_stack-7568bff0848b6ff94a66aff96d074da5.svg.

[17] „Introduction to LangChain,” [Interactiv]. Available: https://python.langchain.com/docs/get_started/introduction.

[18] „User guide: contents — scikit-learn 1.3.2 documentation,” [Interactiv]. Available: https://scikit-learn.org/stable/user_guide.html.

[19] „SciPy documentation — SciPy v1.11.4 Manual,” [Interactiv]. Available: https://docs.scipy.org/doc/scipy/.

[20] „API Reference - Streamlit Docs,” [Interactiv]. Available: https://docs.streamlit.io/library/api-reference.

[21] Brand.md, „Misiunea, atribuţiile şi drepturile BNS,” [Interactiv]. Available: https://statistica.gov.md/ro/misiunea-atributiile-si-drepturile-bns-32.html.

[22] AGENȚIA DE GUVERNARE ELECTRONICĂ, „Portalul de Date,” [Interactiv]. Available: https://date.gov.md/home/about.

[23] „Falcon LLM,” [Interactiv]. Available: https://falconllm.tii.ae/falcon.html.

[24] „MPT,” [Interactiv]. Available: https://www.mosaicml.com/mpt.

[25] „Getting started with Llama 2 - AI at Meta,” [Interactiv]. Available: https://ai.meta.com/llama/get-started/.