



<https://doi.org/10.70728/b.series.fil.v08.i08.067>

КОРПУС МАЪЛУМОТЛАРИНИ ИНТЕЛЛЕКТУАЛ ҚАЙТА ИШЛАШ ТЕХНОЛОГИЯЛАРИ

Тоирова Гули Ибрагимовна

Бухоро давлат университети

*Ўзбек тилишунослиги ва журналистика кафедраси профессори
филология фанлари доктори (DSc)*

tuqulijon@mail.ru

Аннотация. Мазкур мақолада рақамли тилшуносликнинг замонавий йўналишларидан бири сифатида ўзбек тили корпус маълумотларини интеллектуал қайта ишлаш масаласи таҳлил қилинади. Тадқиқотнинг асосий эътибори катта ҳажмдаги тил маълумотларини йиғиш, интеграциялаш, тозалаш, трансформациялаш ва сақлаш жараёнларини Data Engineering ёндашуви асосида ташкил этиш ҳамда уларни Machine Learning технологиялари воситасида интеллектуал таҳлил қилишга қаратилган. Data Engineering, Automated Learning ва Intelligent Automation технологияларининг ўзаро интеграцияси тил маълумотларини қайта ишлашда инсон омилини қисқартириш, маълумотлар сифатини ошириш, таҳлил жараёнини тезлаштириш ва корпусда мавжуд яширин лингвистик қонуниятларни аниқлаш имконини беради. Мақолада Supervised Learning, Unsupervised Learning ва Reinforcement Learning моделларининг корпус маълумотларини қайта ишлашдаги имкониятлари ёритилади. Шунингдек, интеллектуал маълумотлар муҳандислиги ва автоматлаштирилган ўқитишнинг Natural Language Processing билан интеграцияси асосида тил маълумотларидан янги билимларни кашф этишнинг концептуал модели таклиф этилади. Тадқиқот натижасида корпус маълумотларини интеллектуал қайта ишлашни «маълумотларни йиғиш – интеграциялаш – тозалаш – трансформациялаш – машинали ўқитиш – интеллектуал таҳлил – қарор ва билимларни шакллантириш» занжири асосида ташкил этиш мақсадга мувофиқлиги асосланади.

Калит сўзлар: корпус лингвистикаси, компьютер лингвистикаси, Data Engineering, Machine Learning, Automated Learning, Intelligent Automation, Natural Language Processing, NLP, лингвистик маълумотлар, интеллектуал таҳлил, Knowledge Discovery.

Аннотация. В статье анализируется проблема интеллектуальной обработки корпусных данных узбекского языка как одно из современных направлений цифровой лингвистики. Основное внимание уделяется организации процессов сбора,

интеграции, очистки, трансформации и хранения больших объемов языковых данных на основе подхода Data Engineering, а также их интеллектуальному анализу с использованием технологий Machine Learning. Интеграция технологий Data Engineering, Automated Learning и Intelligent Automation позволяет сократить влияние человеческого фактора при обработке языковых данных, повысить качество данных, ускорить аналитические процессы и выявить скрытые лингвистические закономерности, представленные в корпусе. В статье рассматриваются возможности моделей Supervised Learning, Unsupervised Learning и Reinforcement Learning в обработке корпусных данных. Кроме того, на основе интеграции интеллектуальной инженерии данных и автоматизированного обучения с Natural Language Processing предлагается концептуальная модель обнаружения новых знаний на основе языковых данных. В результате исследования обосновывается целесообразность организации интеллектуальной обработки корпусных данных по цепочке «сбор данных – интеграция – очистка – трансформация – машинное обучение – интеллектуальный анализ – формирование решений и знаний».

Ключевые слова: корпусная лингвистика, компьютерная лингвистика, Data Engineering, Machine Learning, Automated Learning, Intelligent Automation, Natural Language Processing, NLP, лингвистические данные, интеллектуальный анализ, Knowledge Discovery.

Abstract. The article analyzes the problem of intelligent processing of Uzbek language corpus data as one of the modern areas of digital linguistics. The main focus is placed on organizing the processes of collecting, integrating, cleaning, transforming, and storing large volumes of linguistic data based on the Data Engineering approach, as well as on their intelligent analysis using Machine Learning technologies. The integration of Data Engineering, Automated Learning, and Intelligent Automation technologies makes it possible to reduce the impact of human intervention in linguistic data processing, improve data quality, accelerate analytical processes, and identify hidden linguistic patterns contained in the corpus. The article examines the potential of Supervised Learning, Unsupervised Learning, and Reinforcement Learning models for corpus data processing. Furthermore, based on the integration of intelligent data engineering and automated learning with Natural Language Processing, a conceptual model for discovering new knowledge from linguistic data is proposed. The study substantiates the feasibility of organizing intelligent corpus data processing according to the following sequence: “data collection – integration – cleaning – transformation – machine learning – intelligent analysis – decision and knowledge formation.”

Keywords: corpus linguistics, computational linguistics, Data Engineering, Machine Learning, Automated Learning, Intelligent Automation, Natural Language Processing, NLP, linguistic data, intelligent analysis, Knowledge Discovery.

КИРИШ.Рақамли технологияларнинг жадал тараққиёти натижасида тилга оид электрон маълумотлар ҳажми кескин ортиб бормоқда. Бугунги кунда электрон китоблар, веб-матнлар, ижтимоий тармоқ материаллари, ахборот ресурслари, рақамлаштирилган архивлар, аудио ва видео материаллардан олинган транскрипциялар тилшунослик тадқиқотлари учун улкан маълумотлар базасини ҳосил қилмоқда. Бундай шароитда тил ҳодисаларини фақат анъанавий, қўлбола таҳлил усуллари асосида ўрганиш катта ҳажмдаги маълумотларни тезкор ва самарали қайта ишлаш имкониятини чеклайди.

Замонавий рақамли муҳитда маълумотларнинг ҳажми, хилма-хиллиги ва янгилашиш тезлигининг ортиши маълумотлар билан ишлашнинг янги технологик ёндашувларини талаб этмоқда. Тақдим этилган асосий тадқиқотда ҳам Data Engineering катта ҳажмдаги маълумотларни йиғиш, қайта ишлаш ва бошқаришни таъминловчи муҳим технологик йўналиш сифатида талқин қилинган. Унда Data Engineering ва автоматлаштирилган ўқитишнинг бирлашиши маълумотларни таҳлил қилиш, прогнозлаш, қарор қабул қилиш ва янги билимларни кашф этиш имкониятларини кенгайтириши таъкидланган.

Мазкур ғояни тилшуносликка татбиқ этганда, асосий масала тил маълумотларини қандай қилиб интеллектуал қайта ишлаш ва улардан янги лингвистик билимларни автоматик ажратиб олиш масаласига бориб тақалади.

Шу нуқтаи назардан, ушбу мақолада Data Engineering ва Automated Learning технологияларининг корпус лингвистикаси билан интеграциялашуви, тил маълумотларини интеллектуал қайта ишлаш босқичлари ҳамда Machine Learning асосида корпусдан янги лингвистик билимларни кашф этиш имкониятлари таҳлил қилинади.

ТАДҚИҚОТ МЕТОДОЛОГИЯСИ.Тадқиқот назарий-таҳлилий характерга эга бўлиб, унда системали таҳлил, қиёсий таҳлил, концептуал моделлаштириш ва технологик жараёнларни функционал таҳлил қилиш методларидан фойдаланилди.

Тадқиқотнинг назарий асосини Data Engineering, Machine Learning, Automated Learning, Intelligent Automation ва Natural Language Processing тушунчаларининг ўзаро муносабати ташкил этади.

Асосий методологик ёндашув маълумотларнинг ҳаётий циклини қуйидаги босқичларда кўриб чиқишга асосланади:

Data Collection → Data Integration → Data Cleaning → Data Transformation → Data Storage → Machine Learning → Intelligent Analysis → Knowledge Discovery.

Манба мақолада Data Engineering маълумотларни йиғиш, қайта ишлаш, трансформациялаш ва сақлаш билан боғлиқ комплекс жараён сифатида берилган. Унинг муҳим вазифалари каторида маълумотларнинг сифати, ишончлилиги, хавфсизлиги ва таҳлилга тайёрлигини таъминлаш кўрсатилган.



1-расм. Ўзбек тили корпус маълумотларини интеллектуал қайта ишлашнинг концептуал модели

1-расмдан кўриниб турганидек, ўзбек тили корпус маълумотларини интеллектуал қайта ишлаш жараёни ўзаро боғлиқ саккиз босқичдан иборат бўлиб, улар Data Engineering, NLP, Machine Learning ва Intelligent Automation технологияларининг интеграцияси асосида шаклланади.

Мазкур методологик занжир тил маълумотларига татбиқ этилганда ҳар бир босқич лингвистик талаблар билан уйғунлаштирилади.

1. DATA ENGINEERING: ТИЛ МАЪЛУМОТЛАРИНИ ЙИГИШ, ИНТЕГРАЦИЯЛАШ ВА ТАЙЁРЛАШ. Data Engineeringнинг асосий вазифаси турли манбалардан келадиган катта ҳажмдаги маълумотларни ягона тизимга келтириш ва уларни кейинги таҳлил учун тайёрлашдан иборат.

Манба тадқиқотда Data Engineering жараёни маълумотларни йиғиш, интеграциялаш, тозалаш ва трансформациялаш каби босқичларга ажратилган. Маълумотларни йиғиш турли базалар, API, файллар, оқимли платформалар ва бошқа манбалардан маълумот олишни қамраб олади.

Корпус лингвистикасида ушбу жараён веб-сайтлар, электрон китоблар, газеталар, журналлар, илмий матнлар, бадий асарлар, расмий ҳужжатлар ва бошқа электрон манбалардан тил материалларини йиғиш билан боғлиқ.

Бироқ турли манбалардан олинган тил маълумотлари бир хил форматга эга бўлмаслиги мумкин. Шу сабабли уларни ягона маълумотлар муҳитига интеграциялаш зарур.

Data Integration жараёнида турли формат ва структурадаги маълумотларни ягона тизимга бирлаштириш амалга оширилади. Манба мақолада бу жараён учун ETL

(Extract, Transform, Load) pipeline, ma'lumotlarni virtuallashtirish va real vaqtda integratsiyalash mexanizmlari kўrsatiladi.

Ўзбек тили корпуси учун бундай ёндашув, айниқса, лотин ва кирилл ёзувларидаги матнларни, турли кодировкалардаги маълумотларни, электрон ҳужжатлар ва веб-материалларни ягона маълумотлар муҳитига бирлаштиришда аҳамиятли.

Бу ерда муҳим жиҳат шундаки, Data Engineering тилшуносни алмаштирмайди. Аксинча, у тилшунос томонидан белгиланган лингвистик талабларни технологик жараёнга айлантириш имконини беради.

Демак:

Лингвистик талаб → маълумотлар модели → автоматик pipeline → стандартлаштирилган тил маълумоти.

Шу муносабат Data Engineering ва тилшунослик ўртасидаги муҳим интеграцион нуқтани ташкил этади.

2. МАЪЛУМОТЛАРНИ ТОЗАЛАШ ВА АВТОМАТЛАШТИРИШ: INTELLIGENT DATA PIPELINE. Корпус маълумотларининг катта ҳажмда бўлиши уларда хатолар, такрорий матнлар, техник белгилар, ностандарт форматлар ва бошқа номутаносибликларнинг пайдо бўлишига олиб келиши мумкин.

Шу сабабли маълумотларни тозалаш интеллектуал корпусни шакллантиришнинг муҳим босқичларидан биридир.

Data Engineeringда Data Cleaning жараёни маълумотлардаги хатоларни аниқлаш, етишмайдиган маълумотларни тўлдириш, номутаносибликларни бартараф этиш ва дубликатларни олиб ташлашни қамраб олади. Бунда deduplication, normalization, imputation ва automated anomaly detection каби усуллардан фойдаланиш мумкин.

Тил маълумотларига нисбатан бу жараён янада мураккаб хусусиятга эга. Чунки тилдаги ҳар қандай фарқни автоматик равишда “хато” деб баҳолаш мумкин эмас. Масалан, сўзнинг вариантлари, диалектал шакллар, услубий фарқлар ёки тарихий лексемалар тилшунослик нуқтаи назаридан қимматли маълумот бўлиши мумкин.

Шунинг учун интеллектуал корпусда техник шовқин ва лингвистик вариативликни бир-биридан фарқлаш муҳим.

Бунда инсон ва интеллектуал алгоритм ўртасидаги вазифалар тақсимои куйидагича бўлиши мумкин:

Босқич	Тилшунос	Интеллектуал тизим
Маълумотни саралаш	Лингвистик мезонни белгилайди	Маълумотни автоматик саралайди
Тозалаш	Хатолар турини белгилайди	Хатоларни автоматик аниқлайди
Нормаллаштириш	Норма ва қоидаи белгилайди	Қоидаи катта ҳажмда қўллайди
Аннотация	Теглар тизимини ишлаб чиқади	Автоматик теглайди
Назорат	Натижани эксперт баҳолайди	Аномал натижани қайта белгилайди

Бу ерда мақоланинг марказий ғояси намоён бўлади: Intelligent Automation инсон меҳнатини бутунлай йўқотиш эмас, балки такрорий ва катта ҳажмдаги жараёнларни

автоматлаштириш орқали инсоннинг илмий-аналитик фаолиятини кучайтиришга хизмат қилади.

Манба мақолада ҳам Intelligent Automation сунъий интеллект, машинали ўқитиш ва жараёнларни оркестрация қилишни бирлаштириши, маълумотларни тайёрлаш, текшириш, тозалаш, трансформациялаш, моделни ўқитиш ва жорий этиш каби жараёнларни автоматлаштириши мумкинлиги таъкидланган.

3. AUTOMATED LEARNING VA KNOWLEDGE DISCOVERY: КОРПУСДАН ЯНГИ БИЛИМЛАРНИ КАШФ ЭТИШ. Data Engineering маълумотларни тайёрлаб берса, Machine Learning улардаги қонуниятларни аниқлашга хизмат қилади.

Манба мақоланинг муҳим жиҳати шундаки, унда автоматлаштирилган ўқитишнинг учта асосий шакли — Supervised Learning, Unsupervised Learning ва Reinforcement Learning ажратилади.

Supervised Learning

Supervised Learning белгиси мавжуд маълумотлар асосида моделни ўқитишга асосланади.

Корпус лингвистикасида у: морфологик теглаш; сўз туркумларини автоматик аниқлаш; матнларни таснифлаш; номланган бирликларни аниқлаш; сентимент таҳлили; лингвистик категорияларни автоматик белгилаш каби вазифаларда қўлланиши мумкин.

Бунда тилшунос томонидан олдиндан белгиланган корпус маълумотлари training data вазифасини бажаради.

Unsupervised Learning эса олдиндан белгиланмаган маълумотлардан яширин структура ва қонуниятларни аниқлашга йўналтирилади.

Манба мақолада унинг асосий усуллари сифатида clustering, dimensionality reduction ва anomaly detection келтирилган.

Корпус лингвистикаси учун бу йўналишнинг аҳамияти катта. Чунки катта ҳажмдаги матнлар орасида олдиндан маълум бўлмаган: лексик гуруҳлар; тематик структуралар; матн типлари; семантик яқинликлар; ноодатий тил ҳодисалари автоматик аниқланиши мумкин.

Айниқса, матн маълумотларини кластерлаш Natural Language Processing, сентимент таҳлили ва тематик моделлаштириш каби вазифаларда қўлланиши мумкинлиги манба мақолада ҳам қайд этилган.

Reinforcement Learningда тизим муҳит билан ўзаро таъсирга киришиб, ҳаракатлари натижасидаги қайта алоқа асосида ўз стратегиясини такомиллаштиради. Манбада бу ёндашув агент, муҳит ва reward signal ўртасидаги муносабат асосида тушунтирилган.

Тил технологияларида бундай ёндашувни автоматик тизимларнинг натижаларини инсон баҳоси ёки бошқа қайта алоқа асосида босқичма-босқич такомиллаштириш концепцияси билан боғлаш мумкин.

4. DATA ENGINEERING + MACHINE LEARNING + NLP: ЯНГИ ЛИНГВИСТИК БИЛИМЛАРНИ КАШФ ЭТИШ МОДЕЛИ. Мақоланинг энг муҳим назарий натижаси — Data Engineering ва Machine Learningни алоҳида технологиялар сифатида эмас, балки ягона интеллектуал маълумотлар цикли сифатида қарашдир.

Бу муносабатни қуйидаги формула билан ифодалаш мумкин:

Data Engineering + Machine Learning + Intelligent Automation = Data-driven Decision Making + Knowledge Discovery

Тилшуносликка татбиқ этилганда мазкур формула қуйидаги кўринишни олади:

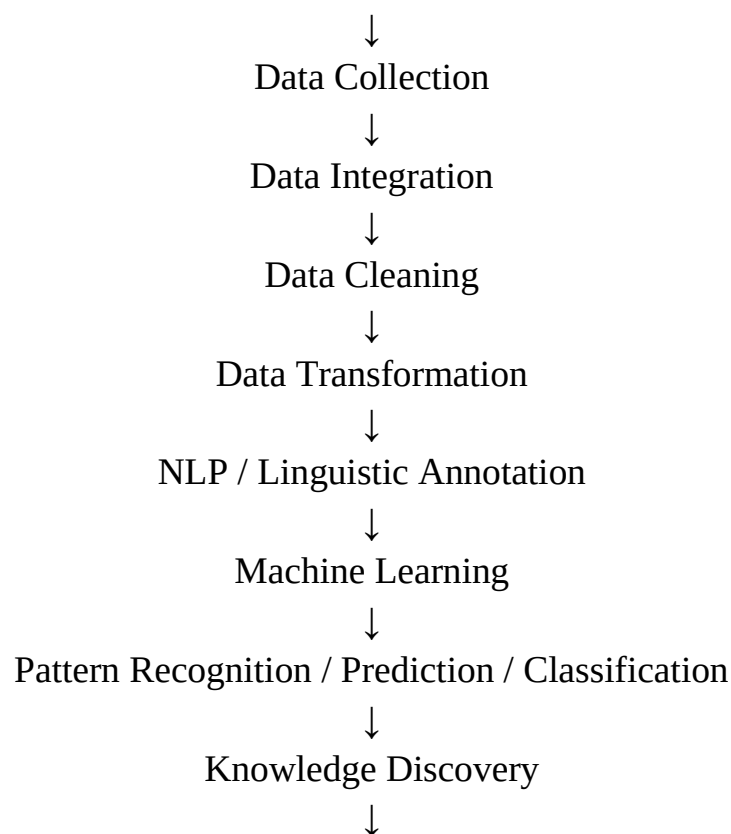
Тил маълумотлари + Data Engineering + NLP + Machine Learning → автоматик лингвистик таҳлил → янги тил билимлари.

Бу ерда Data Engineering маълумотларни тайёрлайди, NLP уларни тил бирликлари нуқтаи назаридан қайта ишлайди, Machine Learning маълумотлардан қонуниятларни ажратади, Intelligent Automation эса ушбу жараёнларни узлуксиз ва масштабланидиган pipelineга бирлаштиради.

Натижада корпус оддий матнлар омборидан интеллектуал тил маълумотлари тизимига айланади.

Ушбу жараёни қуйидаги модельда ифодалаш мумкин:

КАТТА ҲАЖМДАГИ ТИЛ МАЪЛУМОТЛАРИ



ИЛМИЙ ҚАРОР ВА ЯНГИ ЛИНГВИСТИК БИЛИМ

Бу моделда “қарор қабул қилиш” тушунчасини тилшуносликка мос равишда илмий таҳлил қарорлари сифатида талқин қилиш лозим. Яъни тизим тадқиқотчи учун қайси лексик бирликлар кўп қўлланиши, қайси тематик гуруҳлар шаклланиши, қайси

лингвистик бирликлар ўзаро яқинлиги ёки қайси ҳодисалар аномал хусусиятга эгаллигини кўрсатиши мумкин.

НАТИЖАЛАР ВА МУҲОКАМА. Таҳлил натижалари Data Engineering ва Automated Learningнинг интеграцияси катта ҳажмдаги тил маълумотларини қайта ишлашда янги методологик имкониятларни юзага келтиришини кўрсатади.

Биринчидан, Data Engineering корпус маълумотларининг сифатини бошқариш имконини беради. Турли манбалардан келган маълумотларни интеграциялаш, тозалаш, нормаллаштириш ва трансформациялаш кейинги автоматик таҳлилнинг ишончлилиги учун асос бўлади. Манба мақолада ҳам Data Engineering маълумотларни таҳлилга тайёр, аниқ ва изчил ҳолатга келтирувчи асосий инфратузилма сифатида қаралади.

Иккинчидан, Machine Learning инсон томонидан олдиндан кўрсатилмаган қонуниятларни аниқлаш имкониятини кенгайтиради. Айниқса, Unsupervised Learning катта ҳажмдаги белгисиз матн маълумотларидан яширин структураларни излаш учун муҳимдир.

Учинчидан, Intelligent Automation корпус билан ишлашнинг такрорий босқичларини автоматлаштириш имконини беради. Бу маълумотларни қабул қилишдан тортиб, тозалаш, текшириш, моделни ўқитиш ва таҳлил натижаларини янгилашгача бўлган жараёни ягона pipelineга бирлаштириш имконини яратади. Манбада бундай автоматлаштириш маълумотларни катта ҳажмда тезкор қайта ишлаш, тизимларнинг мослашувчанлиги ва реал вақтда таҳлил қилиш имкониятини ошириши таъкидланган.

Тўртинчидан, NLPнинг мазкур тизимдаги ўрни алоҳида аҳамиятга эга. Чунки умумий Data Engineering технологиялари маълумотларни техник жиҳатдан бошқарса, NLP уларнинг лингвистик табиатини ҳисобга олади.

Шунинг учун ўзбек тили корпуси учун энг мақбул ёндашувни:

Data Engineering → NLP → Machine Learning → Intelligent Automation → Knowledge Discovery занжири орқали ифодалаш мумкин.

Бешинчидан, интеллектуал тизимларнинг натижаси ҳар доим инсон томонидан лингвистик экспертиза қилиниши мақсадга мувофиқ. Чунки тил маълумотларидаги вариативлик, контекстуал маъно, полисемия, омонимия, услубий фарқлар каби ҳодисаларни фақат техник алгоритмлар билан тўлиқ изоҳлаш мураккаб.

Шу сабабли мақолада таклиф этилаётган ёндашув “инсонни алгоритм билан алмаштириш” эмас, балки “инсон + интеллектуал тизим” ҳамкорлигига асосланади.

ХУЛОСА. Катта ҳажмдаги маълумотларнинг жадал кўпайиши замонавий тилшунослик олдига тил маълумотларини йиғиш, қайта ишлаш, таҳлил қилиш ва улардан янги билимларни ажратиб олишнинг самарали технологияларини ишлаб чиқиш вазифасини қўймоқда.

Тадқиқот натижалари шуни кўрсатадики, Data Engineering ва Automated Learningни бирлаштириш фақат техник маълумотларни қайта ишлаш жараёни эмас,

балки замонавий тилшуносликнинг рақамли парадигмасини шакллантирувчи муҳим омилдир. Data Engineering маълумотларни йиғиш, интеграциялаш, тозалаш, трансформациялаш ва сақлашни тизимлаштиради, Machine Learning улардаги қонуниятларни автоматик аниқлаш имконини беради. Intelligent Automation эса мазкур жараёнларни ягона, узлуксиз ва масштабланадиган тизимга бирлаштиради.

Шу асосда тил маълумотларини интеллектуал қайта ишлашнинг марказий модели сифатида:

Тил маълумотлари → Data Engineering → NLP → Machine Learning → Intelligent Automation → Knowledge Discovery

занжирини таклиф этиш мумкин.

Мазкур ёндашувда энг муҳим натижа — катта ҳажмдаги маълумотларни шунчаки сақлаш эмас, балки улардан янги лингвистик билимларни автоматик кашф этиш имкониятидир.

Шу билан бирга, интеллектуал технологияларни тилшуносликка жорий этишда маълумотлар сифати, алгоритм натижаларининг ишончлилиги, тизимлар ўртасидаги интеграция, хавфсизлик ва этик масалаларни ҳисобга олиш зарур. Манба тадқиқотнинг хулоса қисмида ҳам айнан маълумотлар сифати, этик масалалар, интероперабеллик ва хавфсизлик интеллектуал Data Engineeringни ривожлантиришдаги асосий муаммолар сифатида кўрсатилади.

Демак, келажакда ўзбек тили корпусларини интеллектуал маълумотлар муҳандислиги, NLP ва автоматлаштирилган ўқитиш технологиялари асосида ривожлантириш корпус лингвистикасини янги сифат босқичига олиб чиқиши, катта ҳажмдаги тил маълумотларидан янги лингвистик қонуниятларни аниқлаш ва илмий қарорларни маълумотларга асосланган ҳолда қабул қилиш имкониятини кенгайтириши мумкин.

ФОЙДАЛАНИЛГАН АДАБИЁТЛАР

1. Biber D., Conrad S., Reppen R. Corpus Linguistics: Investigating Language Structure and Use. – Cambridge: Cambridge University Press, 1998. – 300 p.
2. Bird S., Klein E., Loper E. Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. – Sebastopol, CA: O'Reilly Media, 2009. – 504 p.
3. Elov B. B., Abdisalomova Sh. A., Toirova G. I. The Uzbek Coreference Corpus: Description and Analysis. In: Proceedings of the International Conference on Computer Science and Engineering (UBMK 2025). – Istanbul, Türkiye, 2025. – P. 474–479. DOI: 10.1109/UBMK67458.2025.11206990.
4. Elov B. B., Bekmuradova I. Z., Israilova S. T., Toirova G. I. Tagging Units in the Text and the Bayes Algorithm. In: Proceedings of the IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE 2024). – 2024. – P. 1840–1843. DOI: 10.1109/PIERE62470.2024.10804863.

5. Goodfellow I., Bengio Y., Courville A. Deep Learning. – Cambridge, MA: MIT Press, 2016. – 800 p.
6. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. – Waltham, MA: Morgan Kaufmann, 2012. – 740 p.
7. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning: With Applications in R. 2nd ed. – New York: Springer, 2021. – 607 p.
8. Kleppmann M. Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems. – Sebastopol, CA: O’Reilly Media, 2017. – 616 p.
9. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008. – 506 p.
10. McEnery T., Hardie A. Corpus Linguistics: Method, Theory and Practice. – Cambridge: Cambridge University Press, 2012. – 294 p.
11. Mitchell T. M. Machine Learning. – Boston: McGraw-Hill, 1997. – 414 p.
12. Pradeep A., Rustamov A., Shokirov G., Toirova X., Farkhadovna S. U., Medetovna A. F. Enhancing Data Engineering and Accelerating Learning through Intelligent Automation. In: Proceedings of the 2nd International Conference on Trends in Electrical Electronics and Computer Engineering (TEECCON 2023). – 2023. – P. 104–110.
13. Reis J., Housley M. Fundamentals of Data Engineering: Plan and Build Robust Data Systems. – Sebastopol, CA: O’Reilly Media, 2022. – 450 p.
14. Toirova G. I., Yuldasheva M. R., Elibaeva L. S. Importance of Interface in Creating Corpus. International Journal of Recent Technology and Engineering, 2019, Vol. 8, Issue 2S10. – P. 352–355. DOI: 10.35940/IJRTE.B1057.0982S1019.