

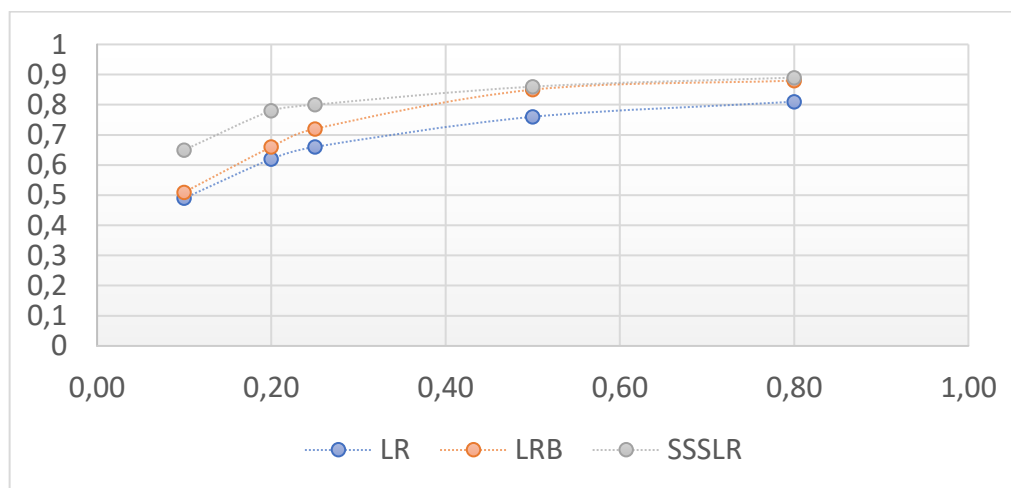


Figure 1. Results of measuring the quality of the classical logistic regression method

The value of 0.80 corresponds to the standard procedure of sliding control on five subsets. It is evident from the graph that the modified logistic regression method (LRB) significantly outperforms the classical method (LR) in situations where the length of the training sample is small compared to the length of the test sample. In a situation where the training sample is of sufficient size, the indicators of the methods are compared.

References

1. Feldman, Ronen, and James Sanger. The textmining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007.
2. Roberta Akemi Sinoara, Joro Antunes and Solange Oliveira Rezende. "Text mining and semantics: a systematic mapping study" Journal of the Brazilian Computer Society (2017) DOI 10.1186/s13173-017-0058-7.
3. Jurek, Anna, Maurice D. Mulvenna, and Yaxin Bi. "Improved lexicon-based sentiment analysis for social media analytics." Security Informatics 4.1 (2015): 1-13.
4. Doshchanova M. Y. et al. Method of selecting informative signs for determining the state of sports readiness of athletes // 2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA). – IEEE, 2023. – C. 1-8.

NLP UCHUN MOS BIG DATA TEXNOLOGIYALARI

Primova Mastura Hakim qizi

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti o'qituvchi
primovamastura@navoiy-uni.uz

Annotatsiya: Zamonaviy Big Data texnologiyalari va vositalari yordami bilan o'zbek tilidagi ulkan matnlar ustida ilgari bajarilishi qiyin bo'lgan vazifalar ham amalga oshirilmoqda. Katta hajmdagi korpuslardan til modellari o'qitish, real vaqt chatlardan ma'noli ma'lumot ajratish, matnlarni semantik indekslash va qidiruv tizimlarini yaratish – bularning barchasi endilikda Hadoop, Spark, NoSQL, Kafka, Elasticsearch singari platformalar integratsiyasi orqali hayotga tatbiq etilmoqda.

Kalit so'zlar: Apache Hadoop, Apache Spark, NoSQL, Apache Kafka, Elasticsearch, bulutli platformalar

ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ, ПОДХОДЯЩИЕ ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

Аннотация: С помощью современных технологий и технологий больших данных также становится сложно реализовать большие объемы текстов на узбекском языке. Обучение языковых моделей из больших корпусов, извлечение содержательной информации из чатов в реальном времени, семантическое индексирование текстов и создание поисковых систем — все это теперь реализуется посредством интеграции таких платформ, как Hadoop, Spark, NoSQL, Kafka и Elasticsearch.

Ключевые слова: Apache Hadoop, Apache Spark, NoSQL, Apache Kafka, Elasticsearch, облачные платформы.

BIG DATA TECHNOLOGIES SUITABLE FOR NATURAL LANGUAGE PROCESSING

Abstract: With the help of modern technologies and big data technologies, it is also becoming difficult to implement large volumes of texts in the Uzbek language. Training language models from large corpora, extracting meaningful information from real-time chats, semantic indexing of texts and creating search engines are all now implemented through the integration of platforms such as Hadoop, Spark, NoSQL, Kafka and Elasticsearch.

Keywords: Apache Hadoop, Apache Spark, NoSQL, Apache Kafka, Elasticsearch, cloud platforms.

Kirish: Matnni qayta ishlash (NLP) va xususan N-gram tahlili katta hajmdagi ma'lumotlarda samarali amalga oshirilishi uchun bir qator zamonaviy Big Data texnologiyalari qo'llaniladi. Quyida ularning asosiy xususiyatlari keltiriladi:

Apache Hadoop: Ochiq kodli katta ma'lumotlarni qayta ishlash platformasi bo'lib, keng **ko'lamli batch (paketli)** rejimda tahlil va saqlashga mo'ljallangan. Hadoop HDFS (Hadoop Distributed File System) orqali terabayt va petabaytlab ma'lumotlarni saqlaydi, MapReduce modeli esa ularni parallel tarzda qayta ishlashni ta'minlaydi[1]. Ushbu tizim yuqori skalalanish va iqtisodiy samaradorlikka ega, ammo real vaqtda ma'lumotni qayta ishlashga mos emas, ya'ni kechikishi katta.

Apache Spark: Umumiy maqsadli, tarqatilgan hisoblash tizimi bo'lib, Hadoop MapReduce'ning cheklovlarini bartaraf etish maqsadida yaratilgan. Spark katta hajmdagi ma'lumotlarda **batch** va **streaming (oqim)** rejimida ishlov berishni, interaktiv analizni hamda *mashinali o'qitish* va *SQL so'rovlarini bajarishni* qo'llab-quvvatlaydi. Operativ xotirada hisoblash va optimallashtirilgan DAG (Directed Acyclic Graph) bajaruv mexanizmi tufayli Spark Hadoop'ga nisbatan ancha tezkor va moslashuvchan hisoblanadi. Biroq uning arxitekturasini nisbatan murakkabroq va katta xotira hamda resurslarni talab qiladi. Spark ekotizimida Spark SQL, Spark Streaming (strukturaviy oqim), MLlib (mashina o'rganish kutubxonasi) kabi komponentlar mavjud bo'lib, katta ma'lumotlarda yuqori darajada tahlil va modellashtirish imkonini beradi.

NoSQL ma'lumotlar bazalari (MongoDB, Cassandra): An'anaviy SQL-relatsion bazalardan farqli o'laroq, NoSQL tizimlari ma'lumotni moslashuvchan, noan'anaviy formatlarda saqlaydi. MongoDB hujjat-oriented baza bo'lib, ma'lumotlarni JSON/BSON hujjatlar shaklida saqlashga ixtisoslashgan va turli yarim tuzilmaviy matnli ma'lumotlarni oson qabul qiladi¹. Apache Cassandra esa keng ustunli (wide-column) taqsimlangan baza bo'lib, tezkor yozish-o'qish amallari va gorizontal skalalanish uchun mo'ljallangan. NoSQL bazalari katta hajmdagi strukturalanmagan matnli ma'lumotlarni shardlash va replikasiya orqali bir nechta serverlarga

¹ <https://aws.amazon.com/ru/compare/the-difference-between-cassandra-and-mongodb>

bo‘lib saqlaydi, shu tariqa gorizontal kengayish va nosozlikka chidamlilikni ta‘minlaydi[2]. Masalan, Cassandra klasteri bir vaqtning o‘zida millionlab yozuvlarni qayd etishi va o‘qishi mumkin, MongoDB esa moslashuvchan sxema va to‘liq indekslash imkoniyati bilan matnli hujjatlarni tezkor qayd etish va qidirishni ta‘minlaydi.

Apache Kafka: Katta tezlikda keluvchi ma‘lumotlar oqimini qayta ishlashga mo‘ljallangan tarqatilgan xabarlar tizimi (messaging platformasi) hisoblanadi. Kafka vositasida real vaqtda ma‘lumotlarni (masalan, loglar, sensor ma‘lumotlari yoki ijtimoiy tarmoq xabarlari) qabul qilish va boshqa tizimlarga uzatish amalga oshiriladi. U publish-subscribe modeli orqali ishlaydi: ishlab chiqaruvchilar (producers) xabarlarni navbatga (topic) yozadi, iste‘molchilar (consumers) esa ularga obuna bo‘lib oladi. Kafka klasteri bir necha broker nodelardan iborat bo‘lib, xabarlarni bo‘laklarga (partitions) ajratib tarqatadi va har bir bo‘lak bir nechta nodlarda replikasiya qilinadi – bu esa baland o‘tkazuvchanlik va past kechikish bilan ishonchli oqim uzatishni ta‘minlaydi. Kafka boshqa Big Data vositalari bilan ham chambarchas integratsiyalasha oladi: masalan, Kafka‘da kelayotgan oqim ma‘lumotlarini Spark Streaming yoki Apache Flink yordamida real vaqtda tahlil qilish, yoki ularni Hadoop HDFS ga yozib borish mumkin.

Elasticsearch: Matnli ma‘lumotlar uchun mo‘ljallangan tarqatilgan qidiruv va tahlil tizimi. Elasticsearch katta hajmdagi matn ma‘lumotlarini indekslash va ularda tezkor qidiruvni amalga oshirishga ixtisoslashgan¹. U JSON hujjatlar tarzida ma‘lumot saqlaydi va avtomatik ravishda matnni tahlil qilib, lug‘aviy birliklarga (tokenlarga) ajratadi hamda invertlangan indeks tuzadi. Bu indeks ma‘lumotlarda so‘zlarning uchrashuv joylarini saqlab, keyinchalik matn bo‘yicha qidiruvni juda tez amalga oshirish imkonini beradi. Elasticsearch klasterlari bir nechta nodelardan tashkil topadi; ma‘lumotlar shardlarga bo‘linib tarqatiladi va har bir shard nusxalari (replica) boshqa nodelarga tarqatiladi – bu yondashuv katta hajmdagi matn korpuslarini gorizontal kengaytirib saqlash va yuqori ishlashni ta‘minlaydi. Elasticsearch yordamida, masalan, katta hajmdagi hujjatlar korpusida bir soniyadan kam vaqt ichida to‘liq matnli qidiruv (near real-time search) amalga oshirish mumkin. Shuningdek, u tabiiy tilni qayta ishlashga oid turli analizatorlar (masalan, so‘zlarni asosiy shaklga keltirish, stop-suzlarni filtrlash) bilan ta‘minlangan.

Bulutli platformalar (AWS, Google Cloud, Azure): Bulutli hisoblash xizmatlari Big Data texnologiyalarini keng miqyosda qo‘llashni ancha soddalashtirdi. AWS, Azure va GCP kabi yirik provayderlar katta hajmdagi ma‘lumotlarni saqlash va qayta ishlash uchun mo‘ljallangan infratuzilmani xizmat sifatida taklif etadilar². Masalan, AWS platformasida Amazon S3 obyektli saqlash (cheksiz skalalanadigan “data lake”), Amazon EMR (Hadoop/Spark klasterlari xizmati), Amazon Kinesis (real vaqt oqimlarini qayta ishlash) va Amazon Comprehend (tayyor NLP xizmati) kabi komponentlar mavjud. Google Cloud platformasida esa Cloud Storage va BigQuery katta ma‘lumotlar ombori, Dataproc (Hadoop/Spark klasterlari), Dataflow (stream va batch ishlov berish) hamda Natural Language API kabi xizmatlar mavjud. Microsoft Azure ham xuddi shunday Azure Blob Storage yoki Data Lake Storage, HDInsight (Hadoop/Spark), Event Hubs (oqimlar uchun) va Cognitive Services (NLP va AI xizmatlari)ni taklif etadi. Bulut texnologiyalari yordamida foydalanuvchilar jismoniy serverlarni boshqarish zaruratisiz istalgan hajmdagi matn ma‘lumotlari bilan ishlovchi klasterlarni ishga tushirishlari, real vaqtda kengaytirish yoki qisqartirishlari mumkin. Xususan, uch yirik provayder katta hajmdagi datasetlar bilan ishlash, mashina o‘rganish va sun‘iy intellekt amaliyotlarini joriy etish uchun kuchli va moslashuvchan muhit taqdim etadi.

¹ <https://www.signitysolutions.com/tech-insights/elasticsearch-real-time-analytics-guide>

² <https://medium.com/@ismahfaris/data-ml-deep-dive-aws-azure-gcp-a32cf470aa1d>

Yuqoridagi texnologiyalar matn ma'lumotlarini qayta ishlashning turli bosqichlarida yoki maqsadlarida qo'llaniladi. Ko'pincha, Hadoop arxitekturasini arxivlangan katta ma'lumotlarni batch tarzda tahlil qilishda, Spark tezkor analitik va iterativ AI vazifalarida, Kafka real vaqtdagi ma'lumot uzatishda, NoSQL bazalari matn va metama'lumotlarni saqlashda, Elasticsearch esa qidiruv va filtrlamada o'zaro to'ldiruvchi tarzda ishlatiladi. Bulut platformalari esa ushbu barcha vositalarni integratsiyalash va orkestratsiya qilish uchun qulay zamin yaratadi.

Adabiyotlar ro'yxati

1. Muller, J. (2020). Scalable Data Architectures for Real-Time Big Data Analytics: A Comparative Study of Hadoop, Spark, and Kafka. *International Journal of AI, BigData, Computational and Management Studies*, 1(4), 8-18.
2. Karimberdiyevich, O. M., Abdulaziz o'g'li, Y. M., & Karimjon o'g'li, M. D. (2024). Big data da matnli ma'lumotlarni o'qitish va tahlil qilish. *IZLANUVCHI*, 1(1), 23-31.
3. <https://aws.amazon.com/ru/compare/the-difference-between-cassandra-and-mongodb>
4. <https://www.signitysolutions.com/tech-insights/elasticsearch-real-time-analytics-guide>
5. <https://medium.com/@ismahfaris/data-ml-deep-dive-aws-azure-gcp-a32cf470aa1d>

ТРАНСФОРМАЦИЯ ТРУДОВЫХ ОТНОШЕНИЙ В УЗБЕКИСТАНЕ: АНАЛИЗ ВНЕДРЕНИЯ УДАЛЕННОЙ РАБОТЫ С ВНЕДРЕНИЕМ ИИ

Атамуратова Гулрух Музафаровна

ассистент кафедры «Общественных наук» филиала КФУ в г. Джизаке
gul.ruh@mail.ru

Аннотация: Статья «Трансформация трудовых отношений в Узбекистане: анализ внедрения удаленной работы с внедрением ИИ» выявляет специфические факторы, влияющие на развитие удаленной работы в Узбекистане, и предлагает практические рекомендации для бизнеса и регуляторов. А также рассматривает оценку потенциала внедрения ИИ для поддержки удаленной работы в Узбекистане и разработку практических рекомендаций.

Ключевые слова: удаленная работа, ИИ, бизнес, государство, работа в Узбекистане, гибридный формат ведения деятельности, Интернет, кибербезопасность.

TRANSFORMATION OF LABOR RELATIONS IN UZBEKISTAN: ANALYSIS OF THE IMPLEMENTATION OF REMOTE WORK WITH THE IMPLEMENTATION OF AI

Annotation: The article "Transformation of Labor Relations in Uzbekistan: Analysis of the Implementation of Remote Work with the Introduction of AI" identifies specific factors influencing the development of remote work in Uzbekistan and offers practical recommendations for businesses and regulators. It also considers the assessment of the potential for the introduction of AI to support remote work in Uzbekistan and the development of practical recommendations.

Keywords: remote work, AI, business, government, work in Uzbekistan, hybrid format of doing business, Internet, cybersecurity.

MEHNAT MUNOSABATLARINING TRANSFORMATSIYASI: SUN'IY INTELLEKTNI JORIY ETGAN HOLDA MASOFAVIY ISHLASHNI TATBIQ ETISH TAHLILI