KAM RESURSLI TILLAR UChUN VEB-SAYTLARDAN MATNLI MA'LUMOTLARNI YIG'ISh

Uteuliev Nietbay Uteulievich – *fizika-matematika fanlari doktori, professor*
Utewlievn@mail.ru
Abdalieva Gulistan Rafikovna – *filologiya fanlari doktori (DSc)*
abdalievagr@gmail.com
Kalmuratov Bekbos Kushkinbaevich – *katta o'qituvchi*
Nukus davlat texnika universiteti
Asenbaeva Dalloris Adilbay qizi – *tayanch doktorant*
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti

СБОР ТЕКСТОВЫХ ДАННЫХ С ВЕБ-САЙТОВ ДЛЯ ЯЗЫКОВ С ОГРАНИЧЕННЫМИ РЕСУРСАМИ

Утеулиев Ниетбай Утеулиевич – *доктор физико-математических наук, профессор*
Абдалиева Гулистан Рафиковна – *доктор филологических наук (DSc)*
Калмуратов Бекбос Кушкинбаевич – *старший преподаватель*
Нукусский государственный технический университет
Асенбаева Даллорис Адилбаевна – *базовый докторант*
Ташкентский университет информационных технологий имени Мухаммада аль-Хоразмий

COLLECTING TEXTUAL DATA FROM WEBSITES FOR LOW-RESOURCE LANGUAGES

Uteuliev Nietbay Uteulievich – *Doctor of Physical and Mathematical Sciences, Professor*
Abdalieva Gulistan Rafikovna – *Doctor of Philological Sciences (DSc)*
Kalmuratov Bekbos Kushkinbaevich – *Senior Lecturer*
Nukus State Technical University
Asenbaeva Dalloris Adilbayevna – *doctoral student*
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi

Tayanch so'zlar: qoraqalpoq tili, web scraping, elektron korpus, kam resursli tillar, NLP.

Rezyume. Maqola kam resursli tillar uchun veb-saytlardan matnli ma'lumotlarni avtomatik yig'ish masalasiga bag'ishlangan bo'lib, qoraqalpoq tili misolida ko'rib chiqiladi. Tadqiqotda Python dasturlash tili, Requests va BeautifulSoup kutubxonalari yordamida kknews.uz, yuz.uz va joqargikenes.uz portallaridan maqola matnlarini yig'ish metodologiyasi ishlab chiqildi. Natijada 1 386 ta navigatsiya sahifasi tahlil qilinib, 18 784 ta maqola va 6 090 553 ta so'zdan iborat birlamchi xom matnlar bazasi shakllantirildi. Ushbu dastlabki xom korpus kelgusida preprocessing bosqichlaridan o'tkazilib, qoraqalpoq tili uchun NLP tizimlari, mashinali tarjima va Transformer modellarini yaratishda boshlang'ich manba sifatida xizmat qilishi mumkin.

Ключевые слова: каракалпакский язык, веб-скрейпинг, электронный корпус, языки с ограниченными ресурсами, NLP.

Резюме. Статья посвящена автоматическому сбору текстовых данных с веб-сайтов для языков с ограниченными ресурсами на примере каракалпакского языка. В исследовании с помощью Python, Requests и BeautifulSoup была разработана методология сбора текстов статей с порталов kknews.uz, yuz.uz и joqargikenes.uz. В результате были проанализированы 1 386 навигационных страниц и сформирована первичная необработанная текстовая база, включающая 18 784 статьи и 6 090 553 слова. Данный первичный необработанный корпус после дальнейшей предварительной обработки может быть использован как начальный ресурс при разработке NLP-систем, машинного перевода и трансформерных моделей для каракалпакского языка.

Key words: Karakalpak language, web scraping, electronic corpus, low-resource languages, NLP.

Summary. This article focuses on the automatic collection of textual data from websites for low-resource languages, using the Karakalpak language as a case study. The study developed a methodology for collecting article texts from kknews.uz, yuz.uz, and joqargikenes.uz using Python, Requests, and BeautifulSoup. As a result, 1 386 navigation pages were analyzed, and a primary raw text database containing 18 784 articles and 6 090 553 words was formed. The collected raw corpus may serve as an initial resource for further preprocessing and for future development of NLP systems, machine translation tools, and Transformer-based models for the Karakalpak language.

Kirish. Zamonaviy sun'iy intellekt va kompyuter lingvistikasi sohasida tabiiy tilni qayta ishlash – Natural Language Processing (NLP) texnologiyalari muhim ilmiy-amaliy yo'nalishlardan biriga aylandi. Matnlarni avtomatik tahlil qilish, mashinali tarjima, savol-javob tizimlari, sentiment tahlili, imlo tekshirish va matn generatsiyasi kabi vazifalar NLP texnologiyalari asosida amalga oshirilmoqda. Ayniqsa, Transformer arxitekturasiga asoslangan BERT kabi modellar ko'plab NLP vazifalarida yuqori samaradorlik ko'rsatmoqda [1, 2]. Biroq bunday modellarni yaratish va sifatli ishlashini ta'minlash katta hajmdagi toza, tartiblangan va reprezentativ matnli ma'lumotlarga bog'liq.

Dunyo tillarining katta qismi raqamli resurslar nuqtai nazaridan kam resursli tillar toifasiga kiradi. Bunday tillar uchun ochiq elektron korpuslar, parallel matnlar, lingvistik ma'lumotlar bazalari, tokenizatorlar va oldindan o'qitilgan til modellari yetarli darajada shakllanmagan. Natijada kam resursli tillarda mashinali tarjima, avtomatik matn tasnifi, axborot qidiruv tizimlari va boshqa NLP ilovalarini ishlab chiqish jarayoni murakkablashadi. Shu sababli kam resursli tillar uchun matnli ma'lumotlarni yig'ish, tozalash va korpus shakliga keltirish dolzarb vazifalardan biri hisoblanadi [3].

Qoraqalpoq tili ham raqamli resurslari yetarli darajada rivojlanmagan tillardan biri hisoblanadi. Mavjud matnli

resurslar koʻpincha veb-saytlar, yangilik portallari va rasmiy sahifalarda tarqoq holda joylashgan boʻlib, ular mashinali oʻqitish algoritmlari uchun tayyor maʼlumotlar toʻplami shaklida emas. Avvalgi tadqiqotlarda Swiss German, Punjabi va boshqa kam resursli tillar uchun internet manbalaridan avtomatik ravishda matn yigʻish orqali elektron korpuslar yaratish mumkinligi koʻrsatilgan [4, 6]. Ushbu maqolaning maqsadi Requests va BeautifulSoup kutubxonalari yordamida qoraqalpoq tilidagi elektron manbalardan maqola matnlarini avtomatik yigʻish va keyingi qayta ishlash bosqichlari uchun birlamchi xom matnlar bazasini shakllantirish metodologiyasini koʻrsatishdan iborat.

Metodologiya. Tadqiqotda qoraqalpoq tili uchun birlamchi xom matnlar bazasini shakllantirish jarayoni maʼlumot manbalarini tanlash, veb-sahifalarni avtomatik yuklab olish, HTML tarkibidan kerakli matnlarni ajratish, birlamchi tozalash va saqlash bosqichlarini oʻz ichiga oladi. Dastlabki xom korpusning leksik va uslubiy xilma-xilligini taʼminlash maqsadida qoraqalpoq tilidagi uchta ishonchli elektron resurs tanlandi: kknews.uz, yuz.uz va joqargikenes.uz. Manbalarni tanlashda ularning ochiq veb-resurs ekanligi, qoraqalpoq tilidagi matnlarni muntazam eʼlon qilishi, tahririy nazoratdan oʻtganligi va turli janrdagi matnlarni qamrab olishi asosiy mezon sifatida belgilandi.

Veb-resurs	Sohasi	Til uslubi	Qamrab olingan sahifalar
kknews.uz	Ommaviy axborot vositalari	Jurnalistik, ijtimoiy-siyosiy	613
yuz.uz	Milliy yangiliklar agentligi	Rasmiy, tahliliy	623
joqargikenes.uz	Davlat boshqaruvi	Qonunchilik, rasmiy-idoraviy	150

Bunday manbalarni tanlash kam resursli tillar uchun internetdan korpus yaratish tajribasiga mos keladi. Masalan, SwissCrawl tizimida ham internetdagi ochiq veb-sahifalar asosiy matn manbasi sifatida ishlatilgan [4].

Maʼlumotlarni yigʻish jarayonida Python dasturlash tili, Requests va BeautifulSoup kutubxonalaridan foydalanildi. Requests veb-sahifalarga HTTP soʻrov yuborish va HTML hujjatni yuklab olish uchun, BeautifulSoup esa yuklab olingan HTML tarkibini tahlil qilish, kerakli teglarni topish, sarlavha, asosiy matn va havolalarni ajratib olish uchun qoʻllanildi. Kahlon va Singh tomonidan Punjabi tili misolida olib borilgan tadqiqotda ham BeautifulSoup statik HTML sahifalardan matnli maʼlumotlarni olishda samarali Python kutubxonasi sifatida taʼriflangan [6].

Ishlab chiqilgan algoritm ikki asosiy bosqichga asoslandi, avval navigatsiya sahifalari koʻrib chiqilib, maqolalarga olib boruvchi havolalar ajratildi, keyin har bir maqola sahifasiga alohida soʻrov yuborilib, uning sarlavhasi va asosiy matni olindi. Jarayon quyidagi ketma-ketlikda amalga oshirildi: 1) sayt URL va sahifalar sonini belgilash; 2) navigatsiya sahifasini Requests orqali yuklab olish; 3) BeautifulSoup yordamida <a> teglarini tahlil qilish; 4) /qr/ yoki /kk/ prefiksli unikal havolalarni ajratish; 5) maqola sahifasidan <h1> va asosiy matn bloklarini olish; 6) matnni filtrlash va korpus fayliga saqlash. Mazkur yondashuv SwissCrawl tizimidagi crawler tamoyiliga yaqin boʻlib, unda ham URL lar roʻyxati olinadi, sahifa yuklanadi, HTML tarkibidan matn va havolalar ajratiladi [4].

BeautifulSoup yordamida sahifaning HTML tuzilmasi obyekt sifatida tahlil qilindi va maqola mazmunini ifodalovchi sarlavha hamda paragraf matnlari ajratib olindi. HTML teglar, menyu elementlari, reklama, navigatsiya bloklari va texnik qismlar korpusga kiritilmadi. Web-korpus yaratishda aynan asosiy matnni ajratib olish muhim bosqich hisoblanadi, chunki sahifalardagi menyu, havola, reklama va boshqa boilerplate qismlar matn sifatini pasaytirishi mumkin [5].

Veb-sahifalardan olingan xom matnlar birlamchi tozalashdan oʻtkazildi, bunda HTML teglar olib tashlandi, ortiqcha boʻsh joylar normallashtirildi, juda qisqa va maʼno bermaydigan paragraflar filtrlandi, matnlar UTF-8 kodirovkasida saqlandi hamda takroriy URL manzillar alohida roʻyxat orqali nazorat qilindi. Serverga ortiqcha yuklama bermaslik uchun soʻrovlar orasida time.sleep yordamida qisqa tanaffuslar qoʻyildi, vaqtinchalik HTTP xatoliklariga chidamlilik uchun esa Retry mexanizmlaridan foydalanildi. Laippala va boshqalar ham web-korpus yaratishda raw HTML matnni tozalash va boilerplate removal bosqichlari matn sifatiga bevosita taʼsir qilishini taʼkidlaydi [5].

Umumiy metodologiyani quyidagicha ifodalash mumkin: manbalarni tanlash → URL sahifalarni aniqlash → HTML sahifani yuklab olish → BeautifulSoup yordamida tahlil qilish → maqola havolalarini ajratish → maqola matnini olish → birlamchi tozalash → saqlash → dastlabki xom matnlar bazasini shakllantirish. Keyingi bosqichlarda ushbu xom maʼlumotlar toʻplamini deduplikatsiya, normalizatsiya, tokenizatsiya va janrlar boʻyicha tasniflash orqali NLP modellarini oʻqitishga tayyor holatga keltirish mumkin.

Natijalar va muhokama. Tadqiqot doirasida ishlab chiqilgan algoritm yordamida uchta 47ortal47n manbadan jami 1 386 ta navigatsiya sahifasi tahlil qilindi. Ushbu sahifalardan 18 784 ta maqola matni ajratib olindi va natijada 6 090 553 ta soʻzdan iborat birlamchi xom matnlar bazasi shakllantirildi. Manbalar boʻyicha yigʻilgan maʼlumotlarning umumiy statistikasi 2-jadvalda keltirilgan.

2-jadval. Manbalar boʻyicha yigʻilgan birlamchi korpus statistikasi

Manba nomi	Sahifalar qamrovi	Maqola-lar soni	Jami soʻzlar	Ulushi (%)
kknews.uz	613	9 808	3 343 302	54.9
yuz.uz	623	7 476	2 561 739	42.1
joqargi kenes.uz	150	1 500	185 512	3.0
JAMI	1 386	18 784	6 090 553	100

Jadvaldan koʻrinib turibdiki, eng 47ort ulush kknews.uz portaliga toʻgʻri keladi: undan 9 808 ta maqola va 3 343 302 ta soʻz yigʻildi. Ikkinchi yirik manba yuz.uz boʻlib, undan 7 476 ta maqola va 2 561 739 ta soʻz ajratib olindi. Joqargikenes.uz hajm jihatidan kichikroq boʻlsa-da, rasmiy-huquqiy va davlat boshqaruviga oid matnlarni oʻz ichiga olgani sababli korpusning uslubiy xilma-xilligini taʼminlashda muhim ahamiyatga ega. Umumiy hisobda har

bir maqolaga o‘rtacha 324 ta so‘z to‘g‘ri keladi, bu esa matnlarning NLP vazifalari uchun foydalanishga yaroqli ekanini ko‘rsatadi.

Yig‘ilgan ma’lumotlarning asosiy afzalligi – ularning turli uslubdagi manbalardan olinganidir. Kknews.uz 48ortal jurnalistik va ijtimoiy-siyosiy uslubdagi matnlarni, yuz.uz rasmiy va tahliliy materiallarni, joqargikenes.uz esa qonunchilik va rasmiy-idoraviy uslubdagi matnlarni qamrab oladi. Bunday xilma-xillik kelajakda qoraqalpoq tili uchun yaratiladigan til modellarining leksik va uslubiy qamrovini kengaytirishga yordam beradi.

Avvalgi tadqiqotlar ham kam resursli tillar uchun internetdan korpus yaratish samarali yo‘l ekanini ko‘rsatadi. Masalan, Swiss German tili uchun yaratilgan SwissCrawl korpusi internetdan avtomatik yig‘ilgan 562 524 ta gap, 62 ming URL va 3 472 domen asosida shakllantirilgan [4]. Ushbu tajriba qoraqalpoq tili kabi raqamli resurslari kam bo‘lgan tillar uchun ham ochiq veb-manbalar asosida korpus yaratish mumkinligini tasdiqlaydi.

Biroq web scraping orqali yig‘ilgan matnlar doimo to‘liq tayyor korpus hisoblanmaydi. Veb-sahifalarda navigatsiya menyulari, reklama bloklari, takroriy havolalar va qisqa texnik matnlar uchrashi mumkin. Shu sababli yig‘ilgan ma’lumotlar dastlabki bosqichda xom korpus sifatida qaraladi. Laippala va boshqalar web-korpus yaratishda HTML sahifalardan olingan matnlarni tozalash, keraksiz qismlarni olib tashlash va asosiy matnni ajratish korpus sifatini oshirishda muhim ekanini ta’kidlaydi [5].

Shakllantirilgan 6 milliondan ortiq so‘zli birlamchi xom korpus qoraqalpoq tili uchun muhim boshlang‘ich resurs hisoblanadi. Ushbu xom ma’lumotlar to‘plami kelgusida maxsus WordPiece yoki BPE tokenizatorlarini yaratish, mavjud ko‘p tilli modellarni qoraqalpoq tiliga moslashtirish, matn tasnifi, axborot qidiruv, mashinali tarjima va terminologik lug‘at tuzish kabi NLP vazifalari uchun dastlabki asos bo‘lib xizmat qilishi mumkin. Shu bilan birga, yirik Transformer modellarini to‘liq o‘qitish uchun yanada kattaroq va chuqur tozalangan korpus talab etiladi.

Xulosa. Ushbu tadqiqotda kam resursli tillar uchun veb-saytlardan matnli ma’lumotlarni avtomatik yig‘ish masalasi qoraqalpoq tili misolida ko‘rib chiqildi. Python dasturlash tili, Requests va BeautifulSoup kutubxonalari yordamida maqola matnlarini yig‘ishga mo‘ljallangan web scraping metodologiyasi ishlab chiqildi. Natijada kknews.uz, yuz.uz va joqargikenes.uz portallaridan 18 784 ta maqola va 6 090 553 ta so‘zdan iborat birlamchi xom matnlar bazasi shakllantirildi.

Tadqiqot natijalari ochiq internet manbalari kam resursli tillar uchun dastlabki matnli ma’lumotlar bazasini yaratishda samarali va amaliy jihatdan qulay manba ekanini ko‘rsatdi. Yangilik portallari va rasmiy veb-saytlardan olingan matnlar qoraqalpoq tilining jurnalistik, ijtimoiy-siyosiy va rasmiy-huquqiy uslublarini qamrab olgani sababli kelajakdagi NLP tadqiqotlari uchun muhim leksik va uslubiy asos yaratadi. Ushbu yondashuv Swiss German va Punjabi kabi boshqa kam resursli tillar bo‘yicha olib borilgan tadqiqotlar tajribasi bilan ham mos keladi [4, 6].

Kelgusida yig‘ilgan xom ma’lumotlar to‘plamini chuqur preprocessing, deduplikatsiya, kodirovka normalizatsiyasi va jumlalarga ajratish bosqichlaridan o‘tkazish, shuningdek uni badiiy adabiyotlar, ilmiy matnlar va ta’limiy resurslar hisobiga kengaytirish maqsadga muvofiq. Xulosa qilib aytganda, veb-saytlardan avtomatik matn yig‘ish qoraqalpoq tilining raqamli resurslarini kengaytirish va uni zamonaviy NLP texnologiyalariga integratsiya qilishda muhim dastlabki bosqich hisoblanadi.

Adabiyotlar

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is All You Need. // Advances in Neural Information Processing Systems, 2017, 30. – P. 5998-6008.

2. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. // arXiv preprint arXiv:1810.04805. 2018.

3. Joshi P., Santy S., Budhiraja R., Bali K., Choudhury M. The State and Fate of Linguistic Diversity in the NLP World. // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. – P. 6282-6293.

4. Linder L., Jungo M., Hennebert J., Musat C., Fischer A. Automatic Creation of Text Corpora for Low-Resource Languages from the Internet: The Case of Swiss German. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020. – P. 2706-2711.

5. Laippala V., Rönnqvist S., Hellström S., Luotolahti J., Repo L., Salmela A., Skantsi V., Pyysalo S. From Web Crawl to Clean Register-Annotated Corpora. Proceedings of the 12th Web as Corpus Workshop, 2020. – P. 14-22.

6. Kahlon N.K., Singh W. Comparative Analysis of Web Scraping Tools for Low-Resource Language Text. // International Journal of Engineering Trends and Technology, 72(1), 2024. – P. 284-299.