UDC 004
SCOPUS CODE 1701
https://doi.org/10.36073/1512-0996-2024-4-90-106

გაუმჯობესებული SEO API-ს შემუშავება ქართული ენისთვის ბუნებრივი ენის დამუშავების მეთოდებით: ტექნოლოგიური მიღწევების გამოყენება ქართულენოვანი ტექსტების შინაარსის ოპტიმიზაციისთვის

ეკატერინე პაპავა ინფორმაციული ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 75
E-mail: papava.e@gtu.ge

თამარ ლომინაძე ინფორმაციული ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 75
E-mail: t.lominadze@gtu.ge

რუსუდან პაპიაშვილი ინფორმაციული ტექნოლოგიების დეპარტამენტი, საქართველოს ტექნიკური უნივერსიტეტი, საქართველო, 0160, თბილისი, მ. კოსტავას 75
E-mail: r.papiashvili@gtu.ge

რეცენზენტები:

მ. ჩხაიძე, სტუ-ის ინფორმატიკისა და მართვის სისტემების ფაკულტეტის პროფესორი
E-mail: m.chkhaidze@gtu.ge

ი. ქართველიშვილი, სტუ-ის ინფორმატიკისა და მართვის სისტემების ფაკულტეტი პროფესორი
E-mail: s.kartvelishvili@gtu.ge

ანოტაცია. სტატია ქართული ენის ფარგლებში ბუნებრივი ენის დამუშავებისა (NLP) და საძიებო სისტემების ოპტიმიზაციის (SEO) კომპლექსურ ურთიერთკვეთას და ამ მოვლენის მნიშვნელობას ეხება. მასში ყურადღება გამახვილებულია იმ უნიკალურ გამოწვევებზე, რომლებიც ქართული ენის რთული მორფოლოგიური შედგენილობით და უნიკალური დამწერლობითაა გამოწვეული, განსაკუთრებით ისეთ სფეროებში, როგორიცაა დასახელებული ერთეულების ამოცნობა (NER), ტოკენიზაცია, ომონიმების ამოცნობა და სხვ. სტატია ხაზს უსვამს მარკირებულ/ანოტირებულ მონაცემთა ნაკრების სიმცირეს ქართულ ენაზე, ბუნებრივი ენის დამუშავების გადაწყვეტილებების აუცილებლობას საძიებო სისტემების ოპტიმიზაციის პერსპექტივიდანაც და აღ-

ნიშნავს იმ უნიკალურ შესაძლებლობებს, რაც ენის გაციფრულებას უნდა მოჰყვეს.

მიუხედავად არაერთი დაბრკოლებებისა, სტატია აგრეთვე ხაზს უსვამს მნიშვნელოვან პოტენციალს, რომელიც მდგომარეობს NLP-ს და ამ ეტაპზე, NER-ის SEO API-ში ჩანერგვაში, ის ითვალისწინებს მომავალს, სადაც ციფრული შინაარსი და მისი პოტენციალი ზედმიწევნით არის შეფასებული დაპროგრამების აპლიკაციის ჩვენ მიერ შემუშავებული ინტერფეისით, რაც მნიშვნელოვნად გააუმჯობესებს მომხმარებლის ძიების საპასუხო მასალას საძიებო სისტემებისგან, მომხმარებლის მიერ ძიებისას ვებგვერდებზე მიღებულ გამოცდილებას და ზოგადად ჩართულობას ქართულ ვებსივრცეში მაღალი ხარისხის მონაცემების საპირწონედ.

სტატია ასევე ეხება საკვანძო სიტყვების კვლევის სპეციფიკას და ქართული ლინგვისტური კონტექსტისთვის დამახასიათებელი შემომავალი ბმულების (backlink) ანალიზის სირთულეებს. საბოლოო ჯამში, ის წარმოადგენს ხელოვნური ინტელექტისა და საძიებო სისტემების ოპტიმიზაციის მეთოდების ინტეგრაციის პოტენციალის კვლევას ტექნოლოგიური ინოვაციებისა და ენობრივი გაგების უწყვეტი შერწყმისას, რაც გზას უხსნის უფრო მდიდარ და ინტერაქტიულ ციფრულ გარემოს ქართული შინაარსისთვის.

საკვანძო სიტყვები: ბუნებრივი ენის დამუშავება; გრამატიკული შეზღუდვები; დასახელებული ერთეულების ამოცნობა; ენის მოდელირება; SEO API, ვებსაიტების ხილვადობა; მანქანური სწავლება საძიებო სისტემების ოპტიმიზაცია; საკვანძო სიტყვების კვლევა, სიტყვის ნაწილის ტეგირება; ტოკენიზაცია; ქართული ენა; შინაარსის ოპტიმიზაცია; შინაარსის ტეგირება.

შესავალი

სტატია იკვლევს ბუნებრივი ენის დამუშავების (NLP) და საძიებო სისტემის ოპტიმიზაციის მეთოდების, ინსტრუმენტების (SEO) კვეთას, სადაც ტექნოლოგიებისა და ლინგვისტიკის თანხვედრა წარმოადგენს როგორც გამოწვევას ისე უდიდეს პოტენციალს ქართული ენისთვის. ხელოვნური ინტელექტის ისეთი მოდელებისა და ბიბლიოთეკების გამოყენებით, მაგალითად, როგორიცაა NLTK და spaCy, ექსპერტები მრავალ ენაზე ენობრივი ბარიერის, ენობრივი ნიუანსებისა და სპეციფიკის წინააღმდეგობებს დიდი ხანია გაუმკლავდნენ, თუმცა, გარკვეული ენები, როგორიცაა ქართული, ცნობილი თავისი რთული ენობრივი მახასიათებლებითა და შეზღუდული გლობალური გამოყენებით, ჯერ კიდევ წარმოადგენს უნიკალურ გამოწვევას. ეს სტატია იკვლევს ქართული ენის სპეციფიკას, განიხილავს NLP-ს ინტეგრაციის სირთულეებს SEO-API -ში და ამ პროცესის აუცილებლობას. ის ხაზს უსვამს ისეთ პრობლემებს, როგორიც შეიძლება არსებობდეს მორფოლოგიურად რთული ენების შემთხვევაში რომლებიც ჯერ კიდევ ტექნიკურ ენად ტრანსფორმირების პროცესის დასაწყისში არიან, აქვთ ვრცელი ანოტირებული მონაცემთა ნაკრების, ენის ეროვნული კორპუსის, სხვადასხვა მონაცემის ციფრული ფორმით დამუშავების ნაკლებობა, შესაბამისად, დეფიციტური მხარდაჭერა საძიებო

სისტემებისგან და განიხილავს ამ დაბრკოლებების გადალახვის ახლებურ და მიზნობრივ ხედვას.

ქართული ენა, ცნობილი თავისი რთული მორფოლოგიური სტრუქტურითა და განსხვავებული დამწერლობით ბუნებრივი ენის დამუშავებისა და განვითარებისთვის საინტერესო, მაგრამ უდიდეს გამოწვევას წარმოადგენს. მიუხედავად იმისა, რომ არაერთი ინსტრუმენტი და მასალა უკვე არსებობს და შემუშავებულია ამ ენისთვის, თითოეული მათგანი ჯერ კიდევ განვითარების ადრეულ ეტაპზეა, განსაკუთრებით აღსანიშნავია ამ ინტეგრაციის არარსებობა საძიებო სისტემების თვალსაზრისითაც, საძიებო სისტემების მხრიდან ამ ენის NLP მხარდაჭერის არარსებობის გამო მომხმარებლის მიერ წამოწყებული ძებნისას SERP-ი (საძიებო სისტემების შედეგების გვერდი) რეალურ შედეგს ვერ იძლევა, რაც განსაკუთრებით დამაბრკოლებელ გარემოებად შეიძლება იქნეს მიჩნეული ქართულ ვებგვერდებზე, ქართულ შინაარსზე მუშაობისას. ეს გარემოება ცხადყოფს, რომ ჯერ კიდევ რჩება სივრცე ქართული NLP ინსტრუმენტების სრული დახვეწისთვის და დიდძალი შრომაა გასაწევი ამ შესაძლებლობების სრულად ათვისებისათვის.

ისეთი მცირედ გამოყენებადი ენები როგორიც ქართულია, მიუხედავად იმისა, რომ წარმოადგენს მორფოლოგიურად მდიდარ, უნიკალურ გრამატიკულ სტრუქტურულ ერთეულს, ბუნებრივი ენის დამუშავების თვალსაზრისით, სწორედ ამ სირთულისა და მრავალფეროვანი ასპექტის გამო ქმნის თავსატეხს. აქ არ ვგულისხმობთ მხოლოდ პროგრამული ნაწილის არარსებობას. აქ უბრალო, ნედლი რესურსების დეფიციტიც, სამუშაო, დაუმუშავებელი რესურსის სიმწირეც უნდა იქნეს ხაზგასმული, რაც თავისთავად იწვევს დახვეწილი, მორგებული ალგორითმების მუშაობის შეფერხებასაც. ნედლი მონაცემების ამოღება, დამუშავება, შემდეგ სასარგებლო წყაროს ინფორმაციად გარდაქმნა არის უპირობოდ აუცილებელი პროცესი, რომელიც მხოლოდ მარტივ მასალად გამოყენებისა და კლასიფიცირების საზღვრებს სცილდება. იგი გულისხმობს ტექსტის სრულყოფილად დამუშავებას ენობრივი ნიუანსების აღსაბეჭდად, რის შედეგადაც უნდა მივიღოთ ორგანიზებული მონაცემთა ნაკრები, რომელიც აუცილებელია ღრმა სწავლებისათვის, ბუნებრივი ენის დამუშავების პროცესისა და შემდეგ ამ მეთოდების ტესტირების აპლიკაციაში ინტეგრირებისათვის. ეს პროცესი ტექნოლოგიური სიზუსტისა და ენობრივი სპეციფიკის ზედმიწევნით სწორ და ერთობლივ მუშაობაზეა გათვლილი.

ანოტირებულ მონაცემთა ნაკრების დასახელებული ერთეულების ამოცნობასთან (Named-entity recognition) მორგება ისეთ გლობალურად ნაკლებად გამოყენებად ენებში როგორიც ქართულია, რთული თავსატეხის ამოხსნას ჰგავს. ის მოიცავს არა მხოლოდ ისეთი მონაცემების იდენტიფიცირებას, როგორიცაა სახელები, ორგანიზაციები და მდებარეობები, არამედ ამ სახელების რთული მორფოსინტაქსური თვისებების გააზრებასაც. გამოწვევები წინადადების სეგმენტიდან სიტყვის ამოცნობამდე და მნიშვნელობის დიფერენციაციამდე მერყეობს, რაც მოითხოვს წესებზე დაფუძნებული და სტატისტიკური მიდგომების შერწყმას.

ძირითადი ნაწილი

ენა სისტემაა, რომელიც წარმოადგენს არა რაღაც ერთეულების მოუწესრიგებელ სიმრავლეს, არამედ, ისეთი ელემენტების ერთობლიობას, რომელთა შორისაც გარკვეული მიმართებები და კავშირები არსებობს [6]. ეს ელემენტები გარკვეულ სტრუქტურებს ქმნის, რომლებიც შეიძლება, ჩვენ, ადამიანებმა, ამოვიცნოთ რაღაც ნაწილის მიხედვით, ვიპოვოთ მონახაზი, დავადგინოთ შეკავშირების ფორმა, კავშირის სახეობა, გამოვიცნოთ ამ სისტემაში რომელიმე ერთეულის შეცვლით გამოწვეული ცვლილება კავშირის სხვა ელემენტებში. ენისთვის დამახასიათებელი სტრუქტურული თვისებრიობა, ენობრივი გამონათქვამების არსებობა, რომლის დროსაც ერთის ფორმა მეორეს განაპირობებს, გამონათქვამის ერთი ფორმის არსებობა კმარა, რომ მეორის ფორმა ვიცოდეთ, მაგალითად, თუ მოცემულია სიტყვა „დაანგრია“, ჩვენ ბუნებრივად დავადგენთ ორ შემადგენელ ნაწილსაც, მაგ. „ქარმა“, „ხუხულა“, კაცმა, სახლი. საქმე სიტყვის მნიშვნელობაში არაა, საქმე მოცემული შემასმენლის ფორმაშია, რომელიც ქვემდებარისგან გარკვეულ მოთხრობით ბრუნვაში დასმას მოითხოვს (ქარმა, კაცმა)...

ბუნებრივი ენის დამუშავების თვალსაზრისით, პროცესი ადამიანის სამეტყველო ენის ციფრულად დამუშავებას გულისხმობს, როდესაც ალგორითმს შეუძლია სწორად დასვას სიტყვების სწორი ფორმები შესაბამის ელემენტებთან, ამოიცნოს აბრევიატურის ან/და შემოკლებული, ნაკლული სიტყვის სრული მნიშვნელობა კონტექსტიდან გამომდინარე, იცნობდეს ენას, როგორც სრულ, ცოცხალ ორგანიზმს და აბრუნებდეს სწორ პასუხებს. მაგალითისთვის, თუ ერთი ელემენტი სახეცვლილი ფორმით, ვთქვათ, „დაანგრევს“ ფორმით შეგვხვდება, ალგორითმმა სახელობითში დასმული ქვემდებარე მიიჩნიოს სწორად, რომ ქარი დაანგრევს და არა ქარმა დაანგრევს.

ენის სპეციფიკისა და ნიუანსების განსაზღვრა ალგორითმისთვის მანქანური სწავლების მეთოდებით უნდა მოხდეს, ლინგვისტური შეზღუდვებისა და გამონაკლისების ხშირად, ინდივიდუალურად განხილვა, ინდივიდუალური ბაზის შექმნა, რაც მნიშვნელოვან როლს შეასრულებს ენის ნიუანსების გაშიფვრაში, რთული მორფემების ამოცნობაში და NER სისტემების სიზუსტის ამაღლებაში[2]. მანქანური სწავლება უფრო მეტია, ვიდრე ტექნოლოგიური გადაწყვეტა; ეს არის ხიდი, რომელიც აკავშირებს სხვადასხვა ენას ციფრულ სფეროსთან.

ამ მოცემულობაში, როგორც ცნობილია, ქართულ ენას საძიებო სისტემების მხრიდან NLP მხარდაჭერა არ აქვს ეს, შეიძლება ვივარაუდოთ, გამოწვეულია ენის სირთულითა და რესურსების არარსებობით, აგრეთვე იმითაც, რომ გლობალურად ეს ენა არ გამოიყენება და, შესაბამისად, დაინტერესებაც არ არსებობს.

ამის მიუხედავად, არაერთი კონტრიბუციის დამსახურებით პროცესი ქართული ენის ტექნიკურად დამუშავების თვალსაზრისით, თითქოს მეტად იხვეწება, გაცილებით უკეთესი მდგომარეობაა საძიებო სისტემებში, ვიდრე ეს, თუნდაც, 2 წლის წინ იყო. ალგორითმი უფრო მგრძნობიარე ხდება ენობრივი დახვეწილობისა და კულტურული კონტექსტების გათვალისწინებითაც, მაგრამ ამის მიუხედავად, შედარებისთვის, საძიებო სისტემების შედეგების ის მონაცემები, რაც ქართულ ენაზე გვხვდება უფრო

პოპულარულ, მაგალითად, ინგლისურენოვან საძიებო სისტემებთან შედარებით, იმდენად განსხვავებული და მწირია, რომ პრაქტიკულად, ვერც ტექნიკური, ვერც ეკონომიკური თვალსაზრისით ვერანაირ შედეგს ვერ დებს. ინტერნეტში საჭირო მასალის მოძიების, მომხმარებლის მიერ პროდუქტის გვერდზე გადასვლის და ბრენდის მიერ სამიზნე მომხმარებელზე წვდომის სარგებელი და მნიშვნელობა იმდენად დიდია, რომ კომპანიები მზად არიან მილიარდი დოლარის ინვესტიცია ჩადონ საძიებო სისტემებში პოზიციების მოპოვებისა და შენარჩუნებისათვის (იხ. სტატისტიკა[1,2]).[1] ამ ფონზე ქართული, ბუნებრივი ენის დამუშავების ალგორითმების SEO API-ებში ინტეგრირება ხსნის ახალ შესაძლებლობებს ქართული ვებგვერდების შესაფასებლად, მათი პოტენციალის განსაზღვრის, კონკურენტული გარემოს შეფასების, წინსვლის, განვითარების, წარმოების სწორად წარმართვისთვის და უზრუნველყოფს მომხმარებელთა უფრო მდიდარ ინტერაქციას, უფრო ზუსტი ძიების შესაძლებლობებს, კონტექსტზე მორგებულ რეკლამას. ასეც რომ არ იყოს, მარტო საკვანძო სიტყვების კვლევის, შინაარსის კონტექსტური გაგებისა და ბექლინკების (შემომავალი ბმულების) ანალიზის ქართული ენის კონტექსტში წარმოება მოითხოვს ტექნოლოგიური ინოვაციებისა და ლინგვისტური გაგების ერთობლიობას.

ქართული ვებსაიტების საძიებო სისტემების ალგორითმების სტანდარტების პერსპექტივიდან შეფასება, შესაბამისობაში მოყვანა და შემდეგ კონკურენტუნარიანობის თვალსაზრისით რეკომენდაციების განსაზღვრა უფრო მეტს მოითხოვს, ვიდრე, უბრალოდ, ცნობილი SEO სტანდარტების დაცვაა, რომელიც აუცილებლად გულისხმობს საძიებო სისტემების ალგორითმებთან გვერდის შინაარსის არქიტექტურისა თუ სხვა მონაცემის ჰარმონიულ თანხვედრას. ის ასევე გულისხმობს კონკრეტული ენის, კონკრეტული სამიზნე აუდიტორიის, ადგილობრივი მოსახლეობის კულტურული ნიუანსებისა და ეროვნული, ენობრივი კონტექსტის ღრმა გააზრებას ტექსტისა და შინაარსის ანალიზში. ეს პროცესი აერთიანებს რაოდენობრივ ანალიზს ერის კულტურული თვისებრიობის გათვალისწინებით და ენის ციფრული ადაპტაციისათვის შინაარსისა თუ სრულად ვებსაიტის შესაძლებლობების შეფასებაში დიდ როლს ასრულებს. ბუნებრივი ენის დამუშავების (NLP) ტექნიკისა და SEO ინსტრუმენტების გაერთიანება აპლიკაციის პროგრამირების ინტერფეისში (API) საშუალებას იძლევა დაინერგოს და მიზნობრივად იქნეს გამოყენებული სისტემური ურთიერთქმედება ტექნოლოგიურ წინსვლასა და ლინგვისტურ გაგებას შორის და ერთიორად გაამდიდროს ციფრული ურთიერთქმედების ლანდშაფტი[4].

ამ ინტეგრაციაში ფუნქციურობა წარმოადგენს საკვანძო მოვლენას ქართული ვებგვერდებისათვის საძიებო სისტემების ოპტიმიზაციის შესაძლებლობების განვითარებისთვის. ის, ასევე, ხდება გზაგასა-

[1] Top 33 ecommerce statistics of 2024 that you should know about. Last Updated: January 17, 2024, [https://www.charle.co.uk/articles/ecommerce-statistics/]

[2] Market share of leading desktop search engines worldwide from January 2015 to July 2023. [https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/]

ყარი, სადაც ენის გატექნოლოგიურება და ციფრული ოპტიმიზაცია სინონიმებად შიძლება განვიხილოთ. ეს თანაკვეთა არ არის მხოლოდ ტექნოლოგიური წინსვლა. იგი წარმოადგენს მრავალფეროვანი ტექნოლოგიების, ალგორითმების, პროგრამებისა და ენობრივი ელემენტების გაერთიანებას. ეს შეკრული ინტეგრაცია ხელს უწყობს უფრო გამდიდრებული და მრავალმხრივი ციფრული ურთიერთქმედების გარემოს შექმნას სხვადასხვა პლატფორმაზე.

ლინგვისტური ლაბირინთის შესწავლა: NLP-ისა და SEO-ს კვეთა

ბუნებრივი ენის დამუშავების (NLP) ინტეგრაცია საძიებო სისტემის ოპტიმიზაციასთან (SEO) არის სინერგიული მიდგომა, რომელიც აერთიანებს ტექნოლოგიურ და ენობრივ გადაწყვეტილებებს. მოწინავე ბიბლიოთეკებმა, როგორიცაა NLTK და spaCy, გაძლიერებული AI მოდელებით, რომლებსაც შეუძლიათ ადამიანის მეტყველების ციფრულ ფორმატში გადატანა, მნიშვნელოვნად გაზარდეს ამ პროცესის მოქნილობა და ეფექტურობა. თუმცა, გამოწვევები რჩება, განსაკუთრებით საძიებო სისტემებში მორფოლოგიურად რთული, მაგრამ ნაკლებად ფართოდ გამოყენებული ენების NLP დამუშავებისას, ამ კონტექსტში ფოკუსს ქართულ ენაზე ვაკეთებთ. ეს ნაკლოვანება აშკარაა ქართულენოვანი ვებსაიტების ტესტირებისას SEO API-ების შესრულებაში. ინსტრუმენტები, როგორიცაა Moz, Ahrefs , SEMrush და სხვა, სავსებით არაეფექტურია გლობალურად ნაკლებად გამოყენებადი ენებისთვის, რადგან ისინი ვერ არის მორგებული გააანალიზონ ამ ენების როგორც უნიკალური, ისე ძირითადი ნიუანსები, რაც გადამწყვეტია ეფექტური ონლაინნავიგაციისთვის.

როგორც უკვე აღვნიშნეთ, ამჟამად, ქართულ ენაზე მორგებული NLP-ინსტრუმენტები განვითარების ახალ ეტაპზეა და მათი სრული პოტენციალი, განსაკუთრებით საძიებო სისტემების ოპტიმიზაციის (SEO) სპეციალიზებულ სფეროში, ჯერ ბოლომდე არ არის რეალიზებული. ამ ინსტრუმენტებში მიმდინარე წინსვლა მნიშვნელოვანი ნაბიჯია ამ ხარვეზის აღმოფხვრისაკენ. თუმცა, ქართულენოვანი ვებგვერდის მრავალი პერსპექტივით გასაანალიზებლად, სრულიად დახვეწილი და საყოველთაოდ მოქმედი ტესტირების ინსტრუმენტების შემუშავების გზა ჯერ კიდევ გასავლელია.

ენები, რომლებიც ხასიათდება მდიდარი მორფოლოგიური სტრუქტურით წარმოადგენენ დიდ სირთულეს ანოტირებული/მარკირებული მონაცემების შემუშავების თვალსაზრისით. კვლევა სცილდება ძირითადი მოდელის ჩამოყალიბების სფეროს; ის მოითხოვს ინტელექტუალური ინსტრუმენტების შექმნას, რომლებიც ადაპტირებული და მგრძნობიარეა ენის ნიუანსების მიმართ. ამ ინსტრუმენტებში ჩვენ უნდა გავითვალისწინოთ ენობრივი სპეციფიკა, სულ მცირე, მაგალითად, რომ დავახასიათოთ, სიტყვის ძირი, ფუძე, აფიქსი, პარადიგმა, მორფემები, გრამატიკული და ლექსიკური მნიშვნელობები, ენობრივი ფორმების უშუალო შემადგენლობები, ამ შემადგენლობებად დაშლის პრინციპები, წინადადებები, წინადადების წევრები, რიცხვი, დრო, კავშირი, გრამატიკულად სწორი და არასწორი ფორმები, აზრიანი და უაზრო გამონათქვამები, ფრაზის გრამატიკული სისწორის განსაზღვრის ფუნქციები და სხვა მრავალი.

NLP-ზე მომუშავე SEO API-ს შექმნა ქართული ენისთვის უმნიშვნელოვანეს გადაწყვეტას წარმოად-

გენს. ამ მცდელობაში ცენტრალური ადგილი უჭირავს ენისა და N-გრამის მოდელების შემუშავებას, ქართული ენის კორპუსის შედგენას, თითოეული სეგმენტი ხელს შეუწყობს ენის ნიუანსების უფრო ფართო გაგებას, რაც არსებითია დახვეწილი მოდელების შემუშავებისათვის.

ნედლი მონაცემების ქმედით ინტელექტად გარდაქმნა ამ პროცესის კრიტიკული ასპექტია. ეს ტრანსფორმაცია სცილდება უბრალო მონაცემთა ანალიზის ფარგლებს და მოიცავს ლექსიკის, როგორც სისტემის აღქმას, მიმართებას ლექსიკურ ერთეულებს შორის, ლექსიკურ სტრუქტურას, ლექსიკურ ომონიმიასა და სინონიმიას, ანტონიმურ წყვილებს, ფრაზეოლოგიზმებს, საკვანძო სიტყვების ძიებას, წინადადების გაყოფას და დასახელებული ერთეულების ამოცნობას. მიზანი არ არის მხოლოდ ენის ტექნიკური ტოკენიზაცია, ის მისი არსის სრულყოფილად აღქმას გულისხმობს. შედეგად უნდა გვქონდეს ზედმიწევნით სტრუქტურირებული მონაცემთა ნაკრები, რომელიც შესაბამისი იქნება ნეირონული ქსელების დასწავლისთვის კომპლექსურ მოთხოვნებზე. მოდელის სწავლებისა და შეფასების თითოეული ეტაპი უნდა ასახავდეს ტექნოლოგიური სიზუსტისა და ენობრივი დახვეწილობის უწყვეტ ინტეგრაციას, რაც მტკიცე საფუძველს ჩაუყრის კომპლექსურ NLP ამოცანებს.

ქართული ენის ინტერპრეტაცია, განსაკუთრებით მორფოსინტაქსურ ანალიზში, არ არის მარტივი ამოცანა. ყოველი სიტყვა და ფრაზა ქართულში წარმოადგენს რთული მორფემებისა და გრამატიკული სტრუქტურების ნაზავს, რომელიც უნიკალურ მნიშვნელობებად და გამოთქმებად ყალიბდება. ამ ენას შეუძლია ერთი სიტყვით გადმოსცეს პიროვნება, საგანი, რაოდენობა, სუბიექტი, ობიექტ(ებ)ი, განწყობა, მიმართულება და სხვა მრავალი ცნება. ამ ნიუანსების გააზრება მოითხოვს ანალიტიკურ მეთოდოლოგიას ღრმა ლინგვისტურ გაგებასთან ერთად, რომელიც ორიენტირებულია არა მხოლოდ სიტყვის იდენტიფიკაციაზე, არამედ ქართულისთვის დამახასიათებელი მდიდარი ენობრივი ელემენტების ყოვლისმომცველ ინტერპრეტაციასა და შეფასებაზე.

ლინგვისტური სირთულეების ამოხსნა: დასახელებული ერთეულის ამოცნობის გამოწვევა ქართული ენისთვის

ლინგვისტურ ტექნოლოგიებში დასახელებული ერთეულების ამოცნობა (NER) კრიტიკული კომპონენტია, ის განსაკუთრებით აქტუალურია ნაკლებად გავრცელებული და ციფრულად ნაკლებად დამუშავებული ენების ანალიზისას, ამ შემთხვევაში ქართული ენისთვის. ეს მოწინავე ტექნიკა ხელს უწყობს დასახელებული ერთეულების კატეგორიზაციას, როგორიცაა საკუთარი სახელები, ორგანიზაციების დასახელებები, გეოგრაფიული დასახელებები და სხვა. ეს ტექნიკა ეფექტურად გარდაქმნის არასტრუქტურირებულ ტექსტს სისტემურად ორგანიზებულ, სამოქმედო ინფორმაციად. NER კარგად არის დანერგილი და ფართოდ გამოიყენება ისეთ გავრცელებულ ენებთან მიმართებით, როგორიც ინგლისურია, მისი დამუშავება ქართული ენისთვის წარმოადგენს გამოწვევების უნიკალურ კომპლექსს, რომელიც გაუგებარი ელემენტების იდენტიფიცირებისთვის პრობლემის გადაჭრის საჭიროების წინაშე დგას და ჯერ კიდევ სავსეა არაერთი სირთულით.

ქართული, თავისი კომპლექსურობიდან გამომდინარე, ერთეული სიტყვის მრავალნაირი ფორმისა და გაგების სირთულე, დიდ გამოწვევებს უქმნის NER სისტემებსაც, რაც მოითხოვს სიზუსტეს და ენის მორფოლოგიური სირთულეების ღრმა გაგებას.

ქართული წინადადების პროგრამულად დაყოფა რთულია. ბევრი ენისგან განსხვავებით, რომელიც პირდაპირ წინადადების დაბოლოებას, ზმნასა ან პუნქტუაციას ეყრდნობა, როგორებიცაა წერტილები, კითხვის ნიშნები ან ძახილის ნიშნები, ქართული ენა თავისი რთული ქვეწყობილი თუ თანწყობილი წინადადებებით ერთგვარ კომპლექსურ და სპეციფიკურ მიდგომას საჭიროებს.

ქართულენოვან მასალაში სათაურს ხშირად აკლია პუნქტუაცია, რაც სიცხადისთვის დამატებით მარკირებას მოითხოვს, წინადადებები, განსაკუთრებით თუ ეს მონაცემები ვებგვერდზეა, გრამატიკის არასრული დაცვითაა დაბეჭდილი. აქ შეიძლება სრულად იგნორირებული იყოს პუნქტუაციის წესები, აკლდეს ასოები, ჰქონდეს უამრავი მრავალწერტილი, აკლდეს მძიმე ან იყოს დასმული შეუსაბამო ადგილას, ეწეროს ერთი ვრცელი აბზაცი.

გარდა ამისა, ერთგვარ ბარიერად შეიძლება, მოგვევლინოს არაერთი პუნქტუაციური წესიც, მაგალითისათვის წერტილები გამოიყენება როგორც აბრევიატურაში, ისე რიცხობრივ გამოსახულებაში, რაც იწვევს პოტენციურ გაურკვევლობას.

მაგ. , ქ. ქუთაისი, მე-5., 2., ა., VI.,

საკითხს კიდევ უფრო ართულებს ქართულ დამწერლობაში დიდი და პატარა ასოს არარსებობა, რაც გამორიცხავს წინადადების დასაწყისზე ჩვეულ მაჩვენებლებს და ამით ალგორითმის მიერ დასაწყისის, დასასრულის ან საკუთარი სახელებისა თუ აბრევიატურების იდენტიფიცირების შესაძლებლობას. ეს ენობრივი განსხვავებულობები მოითხოვს სიტყვების, წინადადებების ზუსტად განსაზღვრის მოწინავე ტექნიკას, დიდ და სწორ მონაცემთა ბაზას, რაც უზრუნველყოფს მნიშვნელოვანი ენობრივი ერთეულების სწორად იდენტიფიცირებას და შემდეგ ალგორითმის მიერ მის დამუშავებას და დასწავლას.

მაგალითად, განვიხილოთ წინადადება: „ჩვენ ხედვად და სმენად უღონონი ვართ და იქ კი... იქ კი რამოდენა ღონე ყოფილა შექმნისა... გიორგი..“ წინადადების შიგნით მრავალწერტილის არსებობა მნიშვნელოვან გამოწვევას უქმნის წინადადებების დაყოფის ტრადიციულ მეთოდოლოგიებს.

უნიკალურ გამოწვევას წარმოადგენს ქართულ ენაში ლექსიკური შესაბამისობის ძიების, სინონიმების იდენტიფიცირების, კონტექსტის შეფასების, ლექსიკონის, თარგმანის პროცესიც. თითოეული სიტყვა ან ტოკენი (წინადადების მცირე ნაწილი, დანაყოფი, პატარა ნიშანი) განხილული უნდა იყოს ინდივიდუალურად, ყველა შესაძლო მორფოსინტაქსური ინტერპრეტაციის გათვალისწინებით, კონტექსტისგან დამოუკიდებლად და შემდეგ, აუცილებლად, კონტექსტში. ლემატიზაციის ამოცანა, ანუ სიტყვების ძირითადი ფორმის მინიჭება, კიდევ უფრო ართულებს ამ ორაზროვნებას. ქართულში სიტყვებს ხშირად მრავალი მნიშვნელობა აქვს, რაც განსაკუთრებით ართულებს მათ ერთმანეთისგან განსხვავებას ან მსგავსების/იდენტური შინაარსის აღმოჩენას.

მაგალითად, ქართული სიტყვა „თვალი“, შეიძლება ნიშნავდეს როგორც „თვალს“, (ინგლ. an eye)

ორგანოს, ასევე „ძვირფას ქვას“ (a gemstone), ნაქსოვის გვირისტს (ინგლ, stitch) ურმის, ეტლის ბორბალს (ინგლ. carriage wheel) და ა.შ. სწორი ინტერპრეტაციის დადგენა კონტექსტური ინფორმაციის არარსებობის შემთხვევაში რთულია, რადგან სიტყვის მნიშვნელობა, დიდწილად, დამოკიდებულია მიმდებარე წინადადების სტრუქტურასა და შინაარსზე.

ქართულში სიტყვის მრავალმნიშვნელოვნება ურთულესი ამოცანაა, რომელიც მოითხოვს სიტყვის სხვადასხვა პოტენციური მორფოსინტაქსური ინტერპრეტაციების გადაწყვეტას მისი კონტექსტური და სემანტიკური გარემოდან გამომდინარე. მიუხედავად იმისა, რომ სტატისტიკური და წესებზე დაფუძნებული ტეგირების მეთოდები იძლევა მრავალმნიშვნელოვნების ამოხსნის გზებს, ქართულის უნიკალური კონტექსტური და სემანტიკური თვისება მოითხოვს ფრთხილად დაბალანსებულ მიდგომას. წესებზე დაფუძნებული ტეგერები, განსაკუთრებით ისინი, რომლებიც იყენებენ გრამატიკულ შეზღუდვებს (CG), გვთავაზობენ ენობრივ გონივრულ გადაწყვეტილებებს. თუმცა, დეტალური, კონტექსტური წესების შემუშავება დროში ინტენსიური პროცესია, რომელიც მოითხოვს ფართო ლინგვისტურ გამოცდილებას და ქართული ენის ნიუანსების ღრმა გააზრებას.

მაგალითი:

განვიხილოთ ქართული ზმნა „იხილე“ (ikhile) რომელიც ორმაგ მნიშვნელობას ატარებს: „დანახვა“ და „თვალის გახელა“. მისი მნიშვნელობა დამოკიდებულია კონტექსტზე. მაგალითად, წინადადებაში „ფერებით შემკული (..) იხილე“ (perebit shemkuli (..) ikile - ფერებით შემკული (..)იხილე), კონტექსტი ნათლად მიუთითებს მნიშვნელობაზე „დანახვა“. პირიქით, „ვიხილები და მოვდივარ“ (vikhilebi da movdivar - თვალებს ვახელ და ძებნას ვიწყებ) იგულისხმება დამალობანას თამაშში გამოყენებული ფრაზა, რომელიც ყველასთვის ნაცნობია ჩვენი ბავშვობიდან.

ქართულ ენაში გრამატიკული შეზღუდვები(CG) ლინგვისტური ანალიზისთვის მყარ ნორმებს გვთავაზობს. ეს ნორმები წესებს ეფუძნება, რომლებიც ვრცელდება სიტყვებზე, ლემაზე, წარმოებულ ლექსიკაზე, მრავალი მნიშვნელობის მქონე სიტყვის ფორმებსა, ტრანსლიტერაციასა და დამარცვლაზე წინადადებების კონტექსტში [5]. ისეთი წესების თანმიმდევრული გამოყენება, ალგორითმში როგორიცაა Select და Remove ხელს უწყობს პოტენციური ალბათობის ზრდას რათა ვიპოვოთ ყველაზე შესაფერისი შესატყვისი [SELECT match-condition [IF context-condition] ; or REMOVE match-condition [IF context-condition] ;]. თუმცა, CG-ითაც კი, სიცხადე ყოველთვის არ არის მიღწეული ენის თვისებრივი სირთულისა და და სიმდიდრის დამსახურებით. სიზუსტესა და სისრულეს შორის ბალანსის მიღწევა მთავარ გამოწვევად რჩება, რომელიც მოითხოვს წესების მუდმივ გაუმჯობესებას და კორექტირებას.

მაგალითი:

"მე" (Me) - "I"

"ვლაპარაკობ" (vlaparakob) - "speak"

"ქართულად" (qartulad) - "in Georgian"

CG ანალიზი

პირის ნაცვალსახელის იდენტიფიკაცია (მე - ME)

SELECT (Pron Pers 1 Nom Sg) ;

ეს წესი ირჩევს წაიკითხოს „მე“, როგორც პირველი პირის ნაცვალსახელი მხოლობით რიცხვში, დასმული სახელობით ბრუნვაში.

ზმნის ფორმის იდენტიფიკაცია (ვლაპარაკობ - vlaparakob)

SELECT (V Act Present 1Sg) IF (-1 Pron Pers 1 Nom Sg) ;

ეს წესი ირჩევს წაიკითხოს ზმნის შემდეგი ფორმა: მოქმედებითი გვარი, აწმყო დრო, პირველი პირის მხოლობითი რიცხვი, თუ მას წინ უძღვის პირველი პირის მხოლობითი რიცხვის ნაცვალსახელი.

ზმნიზედის იდენტიფიცირება - ქართულად (qartulad)

SELECT (Adv) IF (-1 V) ;

ეს წესი ირჩევს „ქართულად“ ზმნიზედად წაიკითხოს, თუ ის ზმნას მოსდევს. სიტყვა მიუთითებს მოქმედების წესსა და რეჟიმზე, მიუთითებს ენაზე, რომლითაც ხდება საუბარი.

პრაქტიკაში სიტყვების არაერთმნიშვნელოვნება ხშირად დამატებით ნაბიჯებსა და დამუშავებას საჭიროებს. სტატისტიკური მეთოდები შეიძლება გამოყენებულ იქნეს ყველაზე სავარაუდო ინტერპრეტაციების პრიორიტეტებისთვის, რაც ქმნის ბალანსს სიზუსტესა და დამახსოვრებულ მონაცემებს შორის. ხელით შესწორებები, რომლებიც შესრულებულია ისეთი ხელსაწყოების გამოყენებით, როგორიცაა Corpuscle, ენათმეცნიერებს საშუალებას აძლევს შეასწორონ შეცდომები პირდაპირ კორპუსში. ეს ნაბიჯები ხაზს უსვამს ქართული ტექსტის დამუშავების კომპლექსურ ხასიათს და ენის როგორც სირთულეების, ისე ანალიზისთვის გამოყენებული გამოთვლითი ინსტრუმენტების გაგების მნიშვნელობას.

ქართული დამწერლობა, უნიკალურია და რთული, განსხვავდება ყველა სხვა დამწერლობისგან, როგორიცაა ლათინური ან კირილიცა, რაც გამოწვევას უქმნის წინასწარ მომზადებული NER-მოდელების ან ხელსაწყოების გამოყენებას, რომლებიც განკუთვნილია უფრო გლობალურად გავრცელებული ენებისთვის. თითოეული სიმბოლო და სიტყვა ქართულში მოითხოვს მორგებულ მიდგომას ეფექტური NER-ისთვის, ენის უნიკალური წმინდა იდენტობის გათვალისწინებით.

ეს მაგალითები ასახავს ქართული ტექსტის ანალიზის რთულ პროცესს, რაც ხაზს უსვამს როგორც ენის, ისე გამოყენებული გამოთვლითი მეთოდოლოგიების დეტალური გაგების აუცილებლობას. ქართულ ენაში ზუსტი ლინგვისტური ანალიზის მიღწევა მოითხოვს წესებზე დაფუძნებულ მეთოდებს, სტატისტიკურ მეთოდებსა და ადამიანურ გამოცდილებას შორის დახვეწილ ურთიერთქმედებას.

ქართულში დასახელებული ერთეულები არა მხოლოდ მრავალფეროვანია, არამედ განსხვავებულია მათი წარმოდგენით, განსაკუთრებით, უცხოურ ტრანსლიტერაციასთან დაკავშირებით. თითოეული ვარიანტი წარმოადგენს განსხვავებულ მნიშვნელობას და ნარატივს. ეფექტური NER-სისტემები ქართულისთვის უნდა იყოს ტექნიკურად მოწინავე და ლინგვისტურად სენსიტიური, რომელსაც შეეძლება ზუსტად ამოიცნოს ეს ნიუანსური განსხვავებები.

მონაცემთა სწორად დამუშავებული ნაკრები გადამწყვეტია მოდელების სწავლების, დახვეწისა და სრულყოფისთვის. ქართული ენისთვის ეს ანოტირებულ მონაცემთა ნაკრების დეფიციტი უფროა,

ვიდრე ტექნიკური დაბრკოლება; ეს არის შეზღუდვა ციფრულ დომენში ენის პოტენციალის სრულად რეალიზაციისთვის.

ამ გამოწვევების საილუსტრაციოდ, ვაჩვენოთ ტოკენიზაცია და NER ქართულად Python's Natural Language Toolkit (NLTK) ბიბლიოთეკის გამოყენებით. ჩვენ მოვახდენთ ქართული წინადადების ტოკენიზაციას და შევეცდებით მასზე დასახელებული ერთეულის ამოცნობას, ტექსტში წარმოგიდგენთ პროცესს და მის შედეგებს.

ტოკენიზაცია და NER-გამოწვევები ქართულად: კერძო შემთხვევის შესწავლა

დასახელებული ერთეულის ამოცნობა (NER) გულისხმობს სხვადასხვა ტიპის სახელების (დასახელებული ერთეულების) ამოცნობას და კლასიფიკაციას, მათ შორის მოიაზრება საკუთარი სახელები, ეთნონიმები, ტოპონიმები, ორგანიზაციისა და ბრენდის სახელები, წიგნებისა და ფილმების სახელები და ა.შ. მრავალი დასახელებული ერთეული შედგება ერთზე მეტი სიტყვისაგან. ევროპულ ენებში დასახელებული ერთეულები, როგორც წესი, იწყება დიდი ასოებით ან სრულად დიდი ასოებისგან შედგება. ეს, რა თქმა უნდა, არ ეხება ქართულს.

დასახელებული ერთეულის ამოცნობის ამოცანის გადაწყვეტისთვის რამდენიმე მიდგომა არსებობს. მათგან უმარტივესი ცნობარების გამოყენებაა – წინასწარ შედგენილი სახელების გაფართოებული, გეოგრაფიული ცნობარებისა, რომლებთან გაცნობაც შესაძლებელია მარკირებისას. უფრო დახვეწილი მეთოდები იყენებს სინტაქსურ და სემანტიკურ გარემოს, რომელშიც ჩვეულებრივ გვხვდება სახელი. ეს ისევ და ისევ, შეიძლება, გაკეთდეს სტატისტიკური და ლინგვისტური, გრამატიკული ტექნიკით.

კვლევაში ძირითადად დავეყრდნობით წინასწარ შედგენილი სახელების გაფართოებული ცნობარის გამოყენების მიდგომას, რომელიც გამყარებული იქნება გრამატიკული შეზღუდვების (CG) წესებით, რომლებიც იღებენ კონტექსტზე დაფუძნებულ პოლისემიურ გადაწყვეტილებებს. გარდა ამისა, ქართული ენის ეროვნულ კორპუსში შესული მრავალი ტექსტი უკვე ხელითაა შევსებული დასახელებული ობიექტებისთვის, ამიტომ აქ განსაკუთრებული სამუშაო არ არის საჭირო.

ქართულ ენაში ტოკენიზაციასთან და დასახელებული ერთეულის ამოცნობასთან (NER) დაკავშირებული გამოწვევების წარმოსაჩენად, განვიხილოთ მაგალითი კონკრეტული ქართული წინადადების ანალიზის საფუძველზე. ჩვენი მიზანია ამ წინადადების ტოკენიზაცია და შემდეგ გამარტივებული NER პროცესის გამოყენება. მნიშვნელოვანია აღინიშნოს, რომ Python-ში Natural Language Toolkit (NLTK), მიუხედავად იმისა, რომ მრავალმხრივი ინსტრუმენტია, არსებითად არ უჭერს მხარს ქართულ NER-ს, რაც მოითხოვს ამ ილუსტრაციისთვის ელემენტარული მიდგომის აუცილებლობას.

გავაანალიზოთ შემდეგი ქართული წინადადება: "ქ. თბილისი არის საქ.-ოს დედაქალაქი."

დავიწყოთ NLTK-დან საჭირო კომპონენტების ჩამოტვირთვით და ფუნქციების იმპორტით:

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
from nltk.tokenize import word_tokenize, sent_tokenize
```

შემდეგ ვახდენთ ტოკენიზაციას, ანუ ცალკეულ ელემენტებად ვაქცევთ:

```
words = word_tokenize(georgian_sentence, language='georgian')
print("Tokenized words:")
print(words)
print()
```

შემდეგ ვცდილობთ გამოვიყენოთ ელემენტარული NER პროცესი. NLTK-ის შეზღუდვების გათვალისწინებით ქართული NER-ის დამუშავებისას, ხელით ვანიჭებთ მდებარეობის ტეგებს ობიექტებს საჩვენებელი მიზნებისთვის:

```
ner_result = [("თბილისი", "LOC"), ("საქართველოს", "LOC")]
print("NER Result:")
for entity, label in ner_result:
    print(f"{entity} - {label}")
```

ამ პროცესის შედეგი ასეთია:

Tokenized words: ['თბილისი', 'არის', 'საქ', 'ოს' 'დიდი', 'ქალაქი', '.']

NER Result:

თბილისი - LOC

ეს გამარტივებული მაგალითი გვიჩვენებს ტოკენიზაციისა და დასახელებული ერთეულის ამოცნობის შესაძლებლობას მარტივი ეტაპებით ქართულ წინადადებაში. თუმცა NLTK-ის შეზღუდული შესაძლებლობები ქართული ენისა და დამწერლობის ნიუანსურ-კომპლექსური პრობლემების გადაჭრაში ხაზს უსვამს უფრო სპეციალიზებული ხელსაწყოებისა და მოდელების საჭიროებას. ქართული ენის ენობრივი სტრუქტურისა და დამწერლობის სირთულე, სრული ანოტირებულ მონაცემთა ნაკრების ნაკლებობასთან ერთად, მნიშვნელოვან გამოწვევებს უქმნის NER-ის ზუსტად შესრულებას [1].

რეალურ სამყაროში, ქართული NER-ის წარმატებული განხორციელება მოითხოვს სპეციალიზებული ალგორითმებისა და მოდელების შემუშავებას, რომლებიც საგანგებოდ შექმნილია ქართული ენის უნიკალური მორფოლოგიური და სინტაქსური მახასიათებლების გაგებისა და დამუშავებისათვის

მანქანური სწავლება, როგორც კატალიზატორი NLP და SEO სინერგიაში

ბუნებრივი ენის დამუშავების (NLP) და საძიებო სისტემების ოპტიმიზაციის (SEO) სინერგიისას მანქანური სწავლება წარმოადგენს გადამწყვეტ ფაქტორს ამ სფეროებში უამრავი თანამდევი გამოწვევის გადასაჭრელად. მორფოლოგიურად კომპლექსური ენის კონტექსტში, ქართული ტექსტების მრავალფეროვან და ვრცელ კორპუსებზე მანქანური სწავლების ალგორითმები მნიშვნელოვან როლს ასრულებს ლინგვისტური სირთულეების გადაჭრაში [7]. ეს ალგორითმები გამოირჩევა რთული მორფემების ამოცნობის, კომპლექსური სიტყვების გარჩევის, გაგებისა და ქართული ენის უნიკალური მახასიათებლების დასწავლა, ადაპტირების მოქნილობით.

წერა (writting)

და-**წერ**-ა (he wrote)

წერილი (a letter)

აღ-**წერ**ა (describe)

წერას ჰყავდა აყვანილი (ან ატანილი) წერამ აიტანა - (Caught in fate's grip)

მაგალითი:

და (sister)

და - კავშირი (and)

მარტივი კლასიფიკაციის მიღმა, მანქანური სწავლების მოდელები ასევე ითვალისწინებს ენობრივ ნიუანსებს და შესაძლებელია დასწავლილ იქნეს ქართული ენის კულტურული და თანამდევი ნიუანსები. SEO აპლიკაციებში, ეს ალგორითმები ხელს შეუწყობს საკვანძო სიტყვების საფუძვლიან კვლევას და გაფართოებულ backlink ანალიზს, დასახელებული ერთეულის ამოცნობის (NER) სისტემების სიზუსტეს მონაცემთა დიდ ნაკრებებში შაბლონების იდენტიფიცირებით.

განვიხილოთ spaCy გამოყენების შემდეგი მაგალითი ქართული NER-ისთვის:

```
import spacy
nlp_georgian = spacy.load("ka_core_news_md")
georgian_sentence = "ბატა არის ფეხსაცმელების მაღაზია, რომელსაც აქვს მაღაზია სითი მოლში და მდებარეობს საქართველოში"
doc_georgian = nlp_georgian(georgian_sentence)
print("Named Entities in the Georgian Sentence:")
for ent in doc_georgian.ents:
    print(ent.text, "-", ent.label_)
```

მოსალოდნელი შედეგი:

დასახელებული ერთეულები ქართულ წინადადებაში:

ბატა - ORG

ფეხსაცმელების მაღაზია - FAC

სითი მოლში - LOC

საქართველოში - LOC

ამ შემთხვევაში, spaCy მოდელი მოსალოდნელია, ადეკვატურად გასაზღვრავდეს ქართულ წინადადებაში დასახელებულ ერთეულებს, კატეგორიებს "ბატას", როგორც ორგანიზაციას (ORG), ფეხსაცმლის მაღაზიას, როგორც ობიექტს (FAC) და ისეთ ადგილებს, როგორიცაა სავაჭრო ცენტრი და "საქართველოში". მნიშვნელოვანია აღინიშნოს, რომ NER-ის ეფექტურობა დამოკიდებულია საბაზისო ენის მოდელის ხარისხსა და მოცულობაზე, ასპექტი, რომელიც იმსახურებს შემდგომ შესწავლას მიმდინარე კვლევაში.

მანქანური სწავლება ამ კონტექსტში სცილდება უბრალო ტექნოლოგიურ ხელსაწყოს როლს [3]. ის წარმოადგენს ენობრივ ხიდს, რომელიც ხელს უწყობს ენების შეუფერხებლად ადაპტირებას ციფრულ ლანდშაფტში. მისი საშუალებით ხდება სტატიაში განხილული ტრანსფორმაციული პოტენციალის აქტუალიზაციაც, რაც ხელახლა განსაზღვრავს ციფრულ ეპოქაში ადამიანის ენისა და ტექნოლოგიების ურთიერთქმედებას და გაგებას.

დასკვნა

ქართულენოვანი ვებკონტენტის შეფასება ჩვეულებრივი მეტრიკის მიღმა

ბუნებრივი ენების დამუშავების (NLP) და საძიებო სისტემების ოპტიმიზაციის (SEO) შერწყმა ქართული ენის კონტექსტში მრავალმხრივ გამოწვევას ბადებს, თუმცა, იმავდროულად, ყოველი გამოწვევა წარმოადგენს ინოვაციური წინსვლის შესაძლებლობას. ტექნოლოგიისა და ლინგვისტიკის ეს შერწყ-

მა რთული პროცესია, რომელშიც ქართული სიტყვების თანდაყოლილი პოლისემიაა, მრავალფეროვანი ნარატივები, რომლებმაც თითოეულ წინადადებას შეიძლება სხვადასხვა მნიშვნელობა მიანიჭოს, თითოეული დასახელებული ერთეულის უნიკალური შინაარსის ძირითადი განმაპირობებელია.

მიუხედავად სირთულეებისა, ბუნებრივი ენის დამუშავების (NLP), თავდაპირველად, დასახელებული ერთეულის ამოცნობის (NER) ინტეგრაცია SEO API-ებში გადამწყვეტი ფაქტორი იქნება ქართულენოვანი, ციფრული შინაარსის დამუშავებისას. ამ სფეროში დეტალური ინფორმაცია, რითაც საძიებო სისტემები აძლიერებს მომხმარებლის ჩართულობას არა მხოლოდ ხელმისაწვდომი, არამედ ზედმიწევნით კატეგორიზებული და სტრუქტურირებული უნდა იყოს. ბუნებრივი ენის დამუშავების მოდელების ინტეგრაცია ძირეულად ზრდის ძიების ფუნქციებს, სთავაზობს უპრეცედენტო დონის სიზუსტეს და კონტექსტურ შესაბამისობას მომხმარებელს მის მიერ წამოწყებულ საძიებო მოთხოვნაზე; რეკლამა გარდაიქმნება კონტექსტურ რეზონანსულ შინაარსად და SEO API-ების მიერ წარმოებული შინაარსის რეკომენდაციები სულ უფრო პერსონალიზებული და ზუსტია, რაც აუმჯობესებს შინაარსზე მუშაობის გამოცდილებას, საძიებო სისტემებში პოზიციონირების კონკურენტუნარიანობას ვებგვერდისთვის და მომხმარებლის საერთო გამოცდილებას.

SEO-ს სფეროში საკვანძო სიტყვების კვლევა ქართული ვებსაიტებისთვის არ არის ადვილი. საკვანძო სიტყვების პოვნა, რომლებიც სრულად ეხმიანება ქართულ კონტექსტს, მოითხოვს ტექნოლოგიური ინოვაციებისა და ლინგვისტური სპეციფიკის დელიკატურ კომბინაციას, იქნება ეს სინონიმების, ომონიმების კვლევა, დასახელებული ერთეულების ამოცნობა, ლოკალიზაცია, ბმების აგება თუ სხვა. თავის მხრივ Backlink (შემომავალი ბმულები) ანალიზიც ტექნიკური პროცედურიდან ცალკეულ, დეტალურ კვლევად უნდა იქნეს განხილული. რომელიც მოითხოვს ალგორითმებს, რომლებსაც შეუძლია ქსელში ნავიგაცია, მაკავშირებელი (anchor) ტექსტების და კონტექსტის იდენტიფიცირება, სისწორისა და მცდარობის შეფასება, პოტენციალის გამოვლენა რაც ქართულენოვანი შინაარსისთვისაა იდენტური.

ქართული ვებსაიტების შეფასება სცილდება ტრადიციულ მეტრიკასა და რაოდენობრივ ანალიზს. ქართულენოვანი საძიებო სისტემების NLP მხარდაჭერა უმნიშვნელოვანესია როგორც ტექნიკური ენის გაციფრულების თვალსაზრისით, ისე მნიშვნელოვნად წინგადადგმულ ნაბიჯია თუნდაც ბიზნესისა და ეკონომიკური გავითარების გადმოსახედიდან.

მეორე მხრივ, ქართული ენა ეროვნული იდენტობისა და სიამაყის სიმბოლოა, განუყოფლად არის დაკავშირებული ქვეყნის მდიდარ ისტორიულ და კულტურულ მემკვიდრეობასთან. შესაბამისად, ქართული შინაარსის ციფრულ წარმოდგენას მნიშვნელოვანი ეროვნული მნიშვნელობაც აქვს. ასეთი კონტენტის ანალიზი და კლასიფიკაცია ტექნიკური შეფასებისა და კულტურული გაგების კომპლექსური ურთიერთკავშირია, რომელშიც ტექნიკური მონაცემები ერწყმის ისტორიულ შეხედულებებს.

NLP და SEO ინსტრუმენტების გაერთიანება ინტეგრირებულ API-ში, ტექნოლოგიური და ლინგვისტური სინერგიის სწორ მაგალითს წარმოადგენს, რის გარეშეც საძიებო სისტემების ოპტიმიზაცია და

პოტენციალის განსაზღვრა ქართულენოვანი კონტენტისთვის, დღესდღეობით, შეუძლებელია. ეს ინტეგრაცია ვერ იქნება უბრალო ტექნოლოგიური მოვლენა, ის აუცილებლად შეიტანს დიდ წვლილს ქართული ენის გაციფრულებაში, რომელიც მდიდარ გამოცდილებას შესძენს ამ პროცესს.

ლიტერატურა

1. Chkaidze, M., Tavdishvili, O,, Chichua, G., & Sophio, B.,(2022) artificial intelligence. ISBN 978-9941-8-2866-9, Educational Scientific Center of Technikal University of Georgia, p. 48.
2. Kachiashvili, Q., (2021) Machine Learning Methods and algoritms. Educational Scientific Cwnter of Technikal University of Georgia, p. 27.
3. Dipanjan, S. (2017), Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems [Book]., O'Reilly Online Learning, Apress.
4. Schwart, V., Schawart, S., & Schwart, M., (2010), The Big Book of NLP, Expanded: 350+ Techniques, Patterns & Strategies of Neuro Linguistic Programming by Shlomo Vaknin. Inner Patch Publishing, 2 Aug.
5. Meurer, P., The Morphosyntactic analysis of Georgian, Uni Research Computing, Bergen, Norway (https://clarino.uib.no/gnc/doc/Morphosyntactic-analysis-of-Georgian.pdf)
6. Gamkrelidze, T., Shaduri I., & Shergelia, N., (2008), Theoretical Linguistics course (second atition) Georgia,
7. Andersen,G., (2012), «Exploring Newspaper Language: Using the Web to Create and Investigate a large corpus of modern Norwegian», Studies in Corpus Linguistics 49, John Benjamins.

UDC 004
SCOPUS CODE 1701
https://doi.org/10.36073/1512-0996-2024-4-90-106

Improved Seo Apl development for Georgian language using natural language processing method: Using tecnilogical advances of obtimize the content of Georgian text

Ekaterine Papava Department of information Tecnology, Georgian Technical University, Georgia, 0160, Tbilisi, 75, M. Kostava str.
E-mail: papava.e@gtu.ge

Tamar Lominadze Department of information Tecnology, Georgian Technical University, Georgia, 0160, Tbilisi, 75, M. Kostava str.
E-mail: t.lominadze@gtu.ge

Rusudan papiashvili Department of information Tecnology, Georgian Technical University, Georgia, 0160, Tbilisi, 75, M. Kostava str.
E-mail: r.papiashvili@gtu.ge

Reviewers:

M. chkaidze, Professor, faculty of informatics and management systems,GTU
E-mail: m.chkhaidze@gtu.ge

I. Qartvelishvili, Professor, faculty of informatics and management systems,GTU
E-mail: s.kartvelishvili@gtu.ge

Abstract. This article delves into the intricate intersection of Natural Language Processing (NLP) and Search Engine Optimization (SEO) specifically tailored for the complexities of the Georgian language. Addressing challenges arising from its intricate morphological structure and unique script, the discussion encompasses Named Entity Recognition (NER), tokenization, homonym recognition, and more. Emphasizing the scarcity of labeled datasets in Georgian, the article underscores the necessity for NLP solutions from an SEO perspective and highlights opportunities brought about by digitizing the language.

Overcoming obstacles, the article underscores the considerable potential within NLP, envisioning a future where our API seamlessly embeds NER in SEO, revolutionizing the evaluation of digital content. This development promises to significantly enhance search response materials, user experiences on websites, and overall engagement within the Georgian digital space, establishing a benchmark for high-quality data.

The content further explores the intricacies of keyword research and the challenges associated with backlink analysis in the unique context of the Georgian language. Ultimately, it serves as an exploration of the potential fusion

of artificial intelligence and SEO methods, creating a seamless synergy of technological innovation and linguistic understanding. This integration aims to cultivate a richer and more interactive digital environment for Georgian content.

Keywords: Content Optimization; Content Tagging; Georgian Language; Grammatical Constraints; Keyword Research; Language Modeling; Machine Learning; Named Entity Recognition; Natural Language Processing; Part-of-Speech Tagging; Search Engine Optimization; SEO API; Tokenization; Website Visibility.

განხილვის თარიღი 04.07.2024
შემოსვლის თარიღი 08.07.2024
ხელმოწერილია დასაბეჭდად 25.12.2024