ქართულენოვანი ტექსტების საწყისი დამუშავების თავისებურებანი NLP ამოცანებისათვის

მაია არჩუაძე, მაგდა ცინცაძე, მანანა ხაჩიძე

maia.archuadze@tsu.ge, magda.tsintsadze@tsu.ge, manana.khachidze@tsu.ge

კომპიუტერული მეცნიერებების დეპარტამენტი
ზუსტ და საბუნებისმეტყველო მეცნიერებათა ფაკულტეტი
ივ. ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი
უნივერსიტეტის ქ. 13, თბილისი

ტექსტის საწყისი დამუშაბება თითქმის ყველა NLP ამოცანის (ტექსტის კომპონენეტების ანალიზი, დაჭდევება, მორფოლოგიური და სინტაქსური ანალიზი, ინფორმაციის ძემნა, მანქანური თარგმნა და ა.შ.) განუყოფელი და აუცილებელი ნაწილია. მის სრულყოფილ და ზუსტ შედეგზეა დამოკიდებული NLP ამოცანების გადაჭრის წარმატებულობა. საყოველთაოდ მიჩნეულია ტექსტის საწყისი დამუშავების ორ ეტაპად გაყოფა -1. ტექსტების სორტირება და 2. ტექსტების სეგმენტაცია. თავის მხრივ ეს ეტაპები მოიცავენ გარკვეულ ქმედებებს, რომლებიც მოითხოვენ ტექსტის სხვადასხვა ბუნებრივენობრივი თავისებურებების გათვალისწინებას, რაც დაკავშირებულია იმაზე თუ რა ენაზეა წარმოდგენილი დასამუშავებელი დოკუმენტი.

ქარლი ენის მორფოლოგიური სირთულე განაპირობებს იმ ფაქტს, რომ სხვადასხვა NLP ამოცანების გადაჭრისას არსებული მეთოდების და ალგორითმების გამოყენება მოდიფიცირების და ზოგჯერ ახლის შემუშავების გარეში შეუძლებელია. ნაშრომში განხილულია NLP ამოცანებუსათვის ტექსტის საწყისი დამუშავების ყველა ძირითადი ბიჯი და ის თავისებურებები, რომლებიც ახლავს ქართულენოვან ტექსტებს. შემოთავაზებულია ალგორითმები, რომლებიც დამუშავებაში ითვალისწინებენ ამ თავისებურებებს და საფუძლად დაედებიან პროგრამულ პაკეტს.

საკვანძო სიტყვები: *ტექსტების დამუშავება, სტოპ სიტყვები*

ლიტერატურა:

1. Kurdi, M.Z. (2016). Natural Language Processing and Computational Linguistics: Speech, Morphology and Syntax, Vol. 1. John
2. Gurusamy, V. and Kannan, S. (2014), 'Preprocessing Techniques for Text Mining', Conference Paper, No. October 2014.
3. M.khachidze, M.Tsintsadze, M. Archuadze, G.Besiashvili. Concept Pattern Based Text Classification System Development for Georgian Text Based Information Retrieval. Baltic J. Modern Computing, Vol. 3 (2015), No. 4, pp. 307–317.
4. M.khachidze, M.Tsintsadze, M. Archuadze. Natural Language Processing (NLP) Based Instrument for Classification of Free Text Medical Records. BioMed Research International. Volume 2016 (2016), Article ID 8313454, 10 pages. http://dx.doi.org/10.1155/2016/8313454