მანქანური სწავლება მომხმარებელთა უკუკავშირის ტექსტების დამუშავებისთვის მარკეტინგის ამოცანებში

გულნარა ჯანელიძე
საინჟინრო მეცნიერებათა დოქტორი, ასოცირებული პროფესორი, საქართველოს ტექნიკური უნივერსიტეტი

ია აფციაური
ინჟინერიის დოქტორი ინფორმატიკაში, ასისტენტ პროფესორი, საქართველოს ტექნიკური უნივერსიტეტი

აბსტრაქტი

თანამედროვე პერიოდში სწრაფად იზრდება მანქანური სწავლების შესაძლებლობები ადამიანების საქმიანობის ყველა სფეროში, მათ შორის, მარკეტინგი შეუძლებელია ტექსტური ინფორმაციის და გამოსახულებების გამოყენებისა და დამუშავების გარეშე. მანქანურმა სწავლებამ მნიშვნელოვნად შეამსუბუქა ვიზუალურ კონტენტთან მუშაობის პროცესი.

მსხვილი კომპანიები ცდილობენ თავიანთი პროდუქტების მიზანმიმართულობასა და პერსონალიზებას. ისინი ამას აკეთებენ ადამიანების ინტერესების გაანალიზებით და მათ მოსაზიდად შესაბამისი მიმართულებით. ეს არის ნაცადი მეთოდი, რომელიც ეხმარება ორგანიზაციებს კონკრეტული აუდიტორიის მოზიდვაში. იმისათვის, რომ მაქსიმალურად იქნას გამოყენებული ინვესტიციები, საჭიროა სწორი ორიენტაცია მყიდველზე. აუდიტორიის სურვილების გაანალიზების გარეშე ფაქტობრივად, ხდება გარისკვა მნიშვნელოვანი დანაკარგებით და მომხმარებელთა უნდობლობით. ნაშრომში მოცემულია კლასტერიზაციის ალგორითმის გამოყენება ტექსტში მსგავსებების გამოსავლენად. წარმოდგენილია კონტექსტზე დამოკიდებული ტექსტის ანალიზის პრობლემები, გრამატიკის საფუძველზე და ასევე, n-გრამებზე დაყრდნობით სახასიათო ნიშნების ამოღების საკითხები. განხილულია საკვანძო ფრაზების ამოღებისა და მთლიანი არსის გამოვლენის ამოცანები.

ნაშრომში წარმოდგენილია მარკეტინგის, კერძოდ გაყიდვების ამოცანებში კლასტერიზაციის ალგორითმების გამოყენების შესაძლებლობები, რომელიც საშუალებას იძლევა მსგავსი თვისების მქონე ადამიანები დაჯგუფდეს გარკვეული პროდუქციის მიმართ მათი ინტერესების მიხედვით. ამავდროულად კლასტერიზაციის ალგორითმი დააჯგუფებს პროდუქტებს, მომხმარებლების უკუკავშირის მიხედვით. შედეგად მიიღება სურათი თუ რამდენად მოთხოვნადია პროდუქტი, რაც შემდგომში გათვალისწინებული იქნება გაყიდვების გასაუმჯობესებლად. ამდენად გაყიდვების ამოცანებში კლიენტების უკუკავშირის ტექსტის დამუშავება შემდგომში მისი კონტენტ-ანალიზისთვის ძალიან ეფექტურია საქონლის შესახებ ინფორმაციის მიღებაში.

საკვანძო სიტყვები: ტექსტში მსგავსებების გამოვლენა, n-გრამების შერჩევა, ტექსტური მონაცემების კლასტერირება.
JEL: M3; C45

DOI: 10.52244/c.2024.11.27

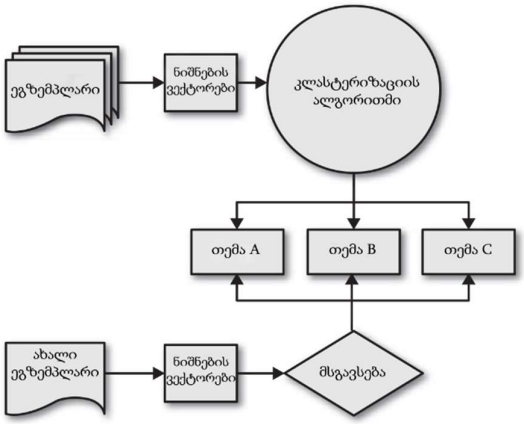
კლასტერირება ტექსტში მსგავსებების გამოსავლენად

კლასტერიზაციის ალგორითმების გამოყენება მეტად ეფექტურია მსგავსი მონაცემების დასაჯგუფებლად. თუ ჩვენი ამოცანა იქნება დავახარისხოთ დიდი რაოდენობის ტექსტური დოკუმენტები, მაგალითად როგორიცაა კულინარიული რეცეპტები, ელექტრონული წერილები ან ჩვენ შემთხვევაში კლიენტების უკუკავშირის ტექსტები, ამისთვის საჭიროა ვცადოთ ყოველი დოკუმენტის წაკითხვა, გამოვყოთ ყველაზე დამახასიათებელი სიტყვები და ფრაზები და დავალაგოთ ისინი დასტებად. თუ რომელიმე დასტა გამოდის ძალიან დიდი შეიძლება ვცადოთ მისი დანაწევრება 2 უფრო პატარა ზომის დასტად. დოკუმენტების განხილვისა და დაჯგუფების შემდეგ, შეგვიძლია ვცადოთ თითოეული დასტის უფრო გულდასმით გამოკვლევა. შევეცადოთ თითოეული დასტიდან საკვანძო სიტყვების ან ფრაზების გამოყოფა, რომ დავახასიათოთ და მივცეთ თითოეულს უნიკალური დასახელება. ფაქტიურად, მსგავსი ტიპის დავალებები ხორციელდება სხვადასხვა სფეროში. ისინი ემყარება ჩვენს უნარს, შევადაროთ დოკუმენტები და დავადგინოთ მათი მსგავსება.

ტექსტური მონაცემების დასაჯგუფებლად მიღებულია კორპუსის ცნება, რომელიც არის ბუნებრივ ენაზე ურთიერთდაკავშირებული დოკუმენტების (ტექსტების) კოლექცია. კორპუსი შეიძლება იყოს დიდი ან პატარა, თუმცა ჩვეულებრივ შედგება რამდენიმე ასეული გიგაბაიტისა და ათასობით დოკუმენტისგან. კორპუსი შეიძლება დაიყოს რამდენიმე კატეგორიის დოკუმენტად ან ცალკეულ დოკუმენტებად. კორპუსში დოკუმენტები შეიძლება განსხვავდებოდეს ზომებით, მაგალითად შეტყობინება ტვიტერიდან ან მთელი წიგნი, ზემოთხსენებული ინფორმაციის ტიპები განსხვავდებიან, თუმცა ყველა მათგანი ტექსტია ზოგჯერ კი მეტაინფორმაცია და წარმოადგენს ერთმანეთთან დაკავშირებულ თემას (How Machine Learning...)

ერთმანეთის მსგავსი დოკუმენტები ერთიანდება ჯგუფებად, რომლებიც კორპუსის შიგნით აღწერენ თემებს და კანონზომიერებებს. ეს კანონზომიერებები შეიძლება იყოს დისკრეტული (მაგალითად, როდესაც ჯგუფები საერთოდ არ კვეთენ ერთმანეთს) ან არაზუსტი (მაგალითად, როდესაც დოკუმენტები იმდენად მსგავსია, რომ მათი გარჩევა ძნელია). ნებისმიერ შემთხვევაში, მიღებული ჯგუფები წარმოადგენენ ყველა დოკუმენტის შინაარსობრივ მოდელს, რომელიც საშუალებას იძლევა ახალი დოკუმენტები მარტივად მივაკუთვნოთ ამა თუ იმ ჯგუფს.

ამჟამად, დოკუმენტების ასეთი დახარისხება ხშირად ხორციელდება ხელით, მაგრამ ეს ამოცანა შესაძლებელია მანქანური სწავლების გამოყენებით (მასწავლებლის გარეშე) უფრო ნაკლებ დროში გადაწყდეს. ამ პროცესში მიზანშეწონილია კლასტერიზაციის ალგორითმის გამოყენება, რომლის მიზანია ნიშნების გამოყენებით გამოავლინოს არამარკირებული მონაცემების დაფარული სტრუქტურა ეგზემპლარის ორგანიზებისთვის არსებითად განსხვავებულ ჯგუფებში.



ნახ. 1. კლასტერირების კონვეიერი

როგორც ნახაზზე ჩანს კონვეიერის გამოსახულებით, კორპუსი გარდაიქმნება ვექტორულ ნიშნებად და გადაეცემა კლასტერიზაციის ალგორითმს ჯგუფის კლასტერების ან თემების განსაზღვრისთვის, მანძილის მეტრიკის გამოყენებით, რომლის თანახმადაც დოკუმენტები, რომლებიც ახლოს არიან ნიშნების სივრცეში ერთმანეთთან, ითვლებიან მსგავსებად. ამის შემდეგ შეიძლება შესრულდეს ახალი დოკუმენტების ვექტორიზაცია და მათი განთავსება უახლოეს კლასტერში (Ali et al., 2016; Bi et al., 2019).

კლასტერირება მსგავსების მიხედვით.

არსებობს მრავალი ნიშანი, რომელსაც შეუძლია დაამოწმოს დოკუმენტების მსგავსება, სიტყვებისა და ფრაზებიდან დაწყებული გრამატიკით და სტრუქტურით დამთავრებული. ჩვენ შეგვეძლო მაგალითად გაგვეერთიანებინა სამედიცინო ჩანაწერები, სადაც აღწერილია პაციენტების მსგავსი სიმპტომები და გვეთქვა, რომ ორი პაციენტი გავს ერთმანეთს *"ლეთარგიითა და გულძმარვით".*

ეფექტური კლასტერიზაციისთვის უნდა განვსაზღვროთ რას ნიშნავს კორპუსიდან ნებისმიერი ორი დოკუმენტისთვის მსგავსება ან განსხვავება. არსებობს მრავალი სხვადასხვა მეტრიკა, რომელთა გამოყენებაც შეიძლება დოკუმენტების მსგავსების დასადგენად; ისინი იდეურად ემყარებიან შესაძლებლობას დოკუმენტები გამოისახოს წერტილების სახით სივრცეში, რომელთა ფარდობითი სიახლოვე განსაზღვრავს მათ მსგავსებას. ამოცანის გადასაწყვეტად მიზანშეწონილია კლასტერიზაციის k-საშუალოს მეთოდის გამოყენება, რომელიც რეალიზებულია NLTK და Scikit-Learn ბიბლიოთეკებში. ეს მეთოდი პოპულარულია მასწავლებლის გარეშე სწავლების ამოცანებისთვის. კლასტერიზაციის ალგორითმი k-საშუალო იწყებს შემთხვევით შერჩეული რაოდენობის k კლასტერებით და ანაწილებს ვექტორიზებულ ეგზემპლარებს კლასტერებად ცენტროიდებთან სიახლოვის მიხედვით, რომლებიც კლასტერების შიგნით ამცირებენ კვადრატების ჯამს.

კონტექსტზე დამოკიდებული ტექსტის ანალიზი

სიტყვებს შორის წარმოქმნილი კონტექსტი, მნიშვნელოვან როლს თამაშობს აზრის გადაცემაში. რაც აუცილებლად გასათვალისწინებელია.

მოდელები, რომლებიც მაგალითად იყენებენ ნორმალიზაციის მეთოდს (არა მარტო), არ ნიშნავს იმას, რომ შეიძლება სრულად უგულებელვყოთ, მართალია ნორმალიზაცია შემდეგ ფრაზებს: *"მას მოსწონს დიდი საყინულე"* და *"მას აქვს დიდი საყინულე"*", გადააქცევს სიტყვების ტომრების იდენტურ ვექტორებად, თუმცა სინამდვილეში ანალიზის საწყის ეტაპზე ეს მოდელები ძალიან სასარგებლოა. ამასთან, არაეფექტური მოდელების ხარისხის გაუმჯობესება ხშირად შესაძლებელია კონტექსტური ნიშნების დამატებით. ერთი უბრალო თუმცა ეფექტური მიდგომა მდგომარეობს გრამატიკების დამატებაში, იმისთვის რომ შეიქმნას შაბლონები, რომლებიც დაგვეხმარება გარკვეულ ფრაზებზე (ისინი მეტ ნიუანსებს გადმოსცემს ვიდრე ცალკეული სიტყვები) აქცენტების გასაკეთებლად.

გრამატიკის საფუძველზე ნიშნების ამოღება

გრამატიკული ნიშნები, როგორიცაა მეტყველების ნაწილები, ენაში დამატებითი ინფორმაციის კოდირების საშუალებას იძლევიან. მოდელის ხარისხის გაუმჯობესების ერთ-ერთ ყველაზე ეფექტურ მეთოდს წარმოადგენს გრამატიკების და პარსერების დანერგვა, შემსუბუქებული სინტაქსური სტრუქტურების შესაქმნელად, დინამიური ტექსტის კოლექციებზე უშუალო ხელშეხებით, რომელთაც შეიძლება ჰქონდეთ დიდი მნიშვნელობა.

იმისათვის, რომ მივიღოთ ინფორმაცია ენაზე რომელზეც არის დაწერილი წინადადება, საჭიროა გრამატიკის წესების ერთობლიობა, რომლებიც განსაზღვრავენ ამ ენაზე ფორმირებული წინადადებების სისწორეს. გრამატიკა ფაქტობრივად წესების ერთობლიობაა, რომლებიც აღწერენ ენის სინტაქსური ერთეულები (წინადადებები, ფრაზები და ა.შ), როგორ უნდა იშლებოდეს ელემენტების შემადგენელ ნაწილებად. განვიხილოთ რამდენიმე მაგალითი ასეთი სინტაქსური კატეგორიების (Bi et al., 2019; Tripathy et al., 2019):

სიმბოლო	სინტაქსური კატეგორია
S	წინადადება (Sentence)
NP	არსებითი სახელის სიტყვათშეთანხმება (Noun Phrase)
VP	ზმნის სიტყვათშეთანხმება (Verb Phrase)
PP	წინდებულიანი სიტყვათშეთანხმება (Prepositional Phrase)
DT	განმსაზღვრელი სიტყვა (Determiner)
N	არსებითი სახელი (Noun)
V	ზმნა (Verb)
ADJ	ზედსართავი სახელი (Adjective)
P	წინდებული (Preposition)
TV	გარდამავალი ზმნა (Transitive Verb)
IV	გარდაუვალი ზმნა (Intransitive Verb)

ნახ. **Error! No text of specified style in document.**.1 სინტაქსური კატეგორიები

კონტექსტისგან თავისუფალი გრამატიკები

გრამატიკების დახმარებით შეიძლება დადგინდეს სხვადასხვა წესები, მეტყველების ნაწილებიდან, ფრაზების ან ფრაგმენტების შეგროვებისთვის. **კონტექსტისგან თავისუფალი გრამატიკა** - ეს არის წესების ერთობლიობა გაერთიანებული სინტაქსური კომპონენტების გააზრებულ სტრიქონებში. მაგალითად, სახელობითი სიტყვათშეთანხმება *"the fridge"* (მაცივარი) მოიცავს განმსაზღვრელ სიტყვას (აღინიშნება ტეგით **DT**, Penn Treebank კრებულიდან) და არსებითი სახელი (**N**). სიტყვათშეთანხმება წინდებულით (**PP**) *"in the fridge"* (მაცივარში) მოიცავს წინდებულს და სახელობით სიტყვათშეთანხმებას (**NP**). სიტყვათშეთანხმება ზმნით (**VP**) *"looks in the fridge"* (მაცივარში იყურება) მოიცავს ზმნას (**V**) და სიტყვათშეთანხმებას წინდებულით (**PP**). წინადადება (**S**) *"Anna looks in the fridge"* (ანა მაცივარში იყურება) მოიცავს საკუთარ სახელს (**NNP**) და კონსტრუქციას ზმნით (**VP**). ამ ტეგების გამოყენებით შესაძლებელია კონტექსტისგან თავისუფალი გრამატიკის დადგენა. NLTK ბიბლიოთეკაში არის ობიექტი nltk.grammar.CFG, რომელიც განსაზღვრავს კონტექსტისგან თავისუფალ გრამატიკას და ურთიერთკავშირებს სხვადასხვა სინტაქსურ ელემენტებს შორის.

სინტაქსური პარსერები

გრამატიკის დადგენის შემდგომ საჭიროა მექანიზმი, რომელიც კორპუსში განახორციელებს გააზრებული სინტაქსური სტრუქტურების სისტემატიურ ძებნას კორპუსში; მექანიზმს, რომელიც ამ როლს თამაშობს ეძახიან პარსერს. გრამატიკა წარმოადგენს ჩვენი ენის კონტექსტში "გააზრებული" ძიების კრიტერიუმს, ხოლო პარსერი ახორციელებს ძებნას. სინტაქსური პარსერი - არის პროგრამული კომპონენტი, რომელიც გარდაქმნის წინადადებებს სინტაქსური ანალიზის ხედ, რომელიც შედგება იერარქიული ელემენტებისგან ან სინტაქსური კატეგორიებისგან.
როცა პარსერს ხვდება წინადადება, ის ამოწმებს მისი სტრუქტურის შესაბამისობას ცნობილ გრამატიკასთან და თუ შეესაბამება, ახორციელებს წინადადების პარსინგს ამ გრამატიკული წესების დაცვით, ანუ წარმოქმნის სინტაქსური ანალიზის ხეს. პარსერებს ხშირად იყენებენ მნიშვნელოვანი სტრუქტურების გამოსავლენად, როგორიცაა სუბიექტის და ობიექტის მოქმედება წინადადებაში, ან სიტყვების თანმიმდევრობა, რომლებიც უნდა იყვნენ დაჯგუფებულნი ყოველ სინტაქსურ კატეგორიაში.

საკვანძო ფრაზების ამოღება.

კორპუსში შესული საკვანძო ტერმინები და ფრაზები, ხშირად იძლევიან შესაძლებლობას მივიღოთ წარმოდგენა ანალიზებადი დოკუმენტების თემაზე ან მათში არსებულ არსებზე. საკვანძო ფრაზების ამოღება მდგომარეობს დინამიური ზომის ფრაზების იდენტიფიცირებასა და გამოყოფაში, იმისთვის, რომ დოკუმენტების თემებში რაც შეიძლება მეტი ნიუანსი იქნეს მოცული.

არსთა ამოღება.

საკვანძო ფრაზების ამოღების ანალოგიურად შეიძლება არსების ამოღების მექანიზმის რეალიზაცია, რომელიც გარდაქმნის დოკუმენტებს "არსების ტომრებად".

ამისთვის გამოდგება ne_chunk უტილიტა NLTK ბიბლიოთეკიდან, რომელიც ქმნის ჩადგმულ ხისებრ სტრუქტურას, შესულს ყოველი წინადადების შემადგენლობაში, სინტაქსური კატეგორიებით და მეტყველების ნაწილის ტეგებით.

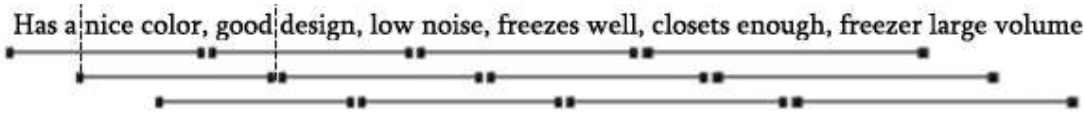
n-გრამებზე დაყრდნობით ნიშნების ამოღება

სამწუხაროდ, გრამატიკაზე დაფუძნებული მიდგომები ყოველთვის არ იძლევიან კარგ შედეგებს. პირველ რიგში ისინი გარკვეულ წილად დამოკიდებულნი არიან მეტყველების ნაწილის ტეგებით სიტყვათა წარმატებულ მარკირებაზე, ანუ უნდა ვიყოთ დარწმუნებულები, რომ ჩვენი მექანიზმი სწორად ადგენს არსებით სახელებს, ზმნებს, ზედსართავ სახელებს და სხვა მეტყველების ნაწილებს. მეტყველების ნაწილის მარკირების მექანიზმი თავდაპირველად default-ით არის განსაზღვრული, ასეთი მარკირება არასტანდარტული ან დაუხვეწავი ტექსტის ანალიზის დროს მარტივად უშვებს შეცდომებს.

გრამატიკაზე დაფუძნებით თვისებების ამოღება ცოტათი მოუქნელია იმის გამო, რომ საჭიროა გრამატიკის წინასწარი განსაზღვრა. თუმცა ხშირად ძალიან ძნელია წინასწარ გაგება რომელი გრამატიკული შაბლონი შეგვაძლებინებს ტექსტში ტერმინების და ფრაზების წარმოჩენას ყველაზე ეფექტურად.

განვიხილოთ ამ პრობლემის გადაჭრა n-გრამების მეშვეობით, რომელიც იძლევა ლეკსემების თანმიმდევრობის იდენტიკაციის უფრო განზოგადებულ მეთოდს.

განვიხილოთ წინადადება *"Has a nice color, good design, low noise, freezes well, closets enough, freezer large volume."* (აქვს კარგი ფერი, კარგი დიზაინი, ნაკლები ხმაურიანია, ყინავს კარგად, განყოფილებები საკმარისია, საყინულე დიდი ზომისაა). n ფიქსირებული სიგანით მთლიანი მიმდევრობის სკანირებისას, შესაძლებელია შევაგროვოთ ყველა შესაძლო უწყვეტი მიმდევრობის ლეკსემა. **უნიგრამი** არის n-გრამი სადაც n = 1, **ბიგრამი** სადაც n = 2, **ტრიგრამი** - n = 3, და ასე შემდეგ. მაგალითად ბიგრამია შემდეგი ლეკსემების კორტეჟები: (*"nice ", " color"*) ან (*"good", "design"*), ტრიგრამების შესაბამისად სამ ელემენტიანი კორტეჟი, **ტეტრაგრამის** ოთხი და ა.შ. ნახ. 2-ზე ნაჩვენებია წინადადების დაყოფის ტრიგრამების მიმდევრობა.



ნახ. 2. წინადადების დაყოფა ფიქსირებული n-ით

რომ გამოვყოთ ტექსტიდან ყველა n-გრამი, საკმარისია გავაანალიზოთ სიტყვების სია გარკვეული ფიქსირებული სიგანით. რასაც Python-ზე ასე გავაკეთებთ:

```
def ngrams(words, n = 2):
    for idx in range(len(words)-n+1):
        yield tuple(words[idx:idx+n])
```

რომ გამოვიყენოთ ეს ფუნქცია ზემოთ აღნიშნულ წინადადებაზე მივიღებთ:

```
words = [
    "Has", "a", "nice", "color", ",", "good", "design", ",", "low", "noise", ",", "freezes", "well",
    ",", "closets", "enough", ",", "freezer", "large", "volume", ".",]
```

```
for ngram in ngrams(words, n = 3):
    print(ngram)
```

```
('Has','a','nice')
('a','nice','color')
('nice','color',',')
('color',',','good')
(',','good','design')
('good','design',',')
('design',',','low')
(',','low','noise')
('low','noise',',')
('noise',',','freezes')
(',','freezes','well')
('freezes','well',',')
('well',',','closets')
(',','closets','enough')
('closets','enough',',')
('enough',',','freezer')
(',','freezer','large')
('freezer','large','volume')
('large','volume','.')
```

n-გრამის ზომის შერჩევა

n = 2 ზომის არჩევისას ვღებულობთ *" Has a", "good design"* და *"low noise"*. თუმცა ეს მოდელი არაა იდეალური, ის წარმატებულად ადგენს სამ შესაბამის არსს დიდი გამოთვლითი დანახარჯების გარეშე. მეორეს მხრივ, ამ ზომის მოდელს (n = 2) არ აქვს შესაძლებლობა დაიჭიროს ტექსტის გარკვეული ნიუანსები. მაგალითად, თუ ტექსტში მოცემულ წინადადებაში აღნიშნულია რამდენიმე მახასიათებელი, მაშინ ბიგრამი " Has a" შეიძლება არაერთმნიშვნელოვანი აღმოჩნდეს. იმისთვის, რომ ყველა სიტყვათშეთანხმება მოვიცვათ, " Has a nice color", საჭიროა ავირჩიოთ n = 4:

```
('Has', 'a', 'nice', 'color')
('a', 'nice', 'color', ',')
('nice', 'color', ',', 'good')
('color', ',', 'good', 'design')
(',', 'good', 'design', ',')
('good', 'design', ',', 'low')
('design', ',', 'low', 'noise')
(',', 'low', 'noise', ',')
('low', 'noise', ',', 'freezes')
('noise', ',', 'freezes', 'well')
(',', 'freezes', 'well', ',')
('freezes', 'well', ',', 'closets')
('well', ',', 'closets', 'enough')
(',', 'closets', 'enough', ',')
('closets', 'enough', ',', 'freezer')
('enough', ',', 'freezer', 'large')
(',', 'freezer', 'large', 'volume')
('freezer', 'large', 'volume', '.')
```

შედეგებიდან გამომდინარე შეგვიძლია დავასკვნათ, რომ დიდი ზომის n-გრამის შედეგები მოცული იქნება დუბლირებული არსებით. ეს გაართულებს ალბათობების განსაზღვრას, რითიც ვსაზღვრავდით ანალიზის მიზნებს. ამის გარდა, n-ის ზრდით იზრდება შესაძლო სწორი n-გრამების რიცხვიც, რის გამოც მცირდება ალბათობა კორპუსში ყველა სწორი n-გრამის ამოცნობისა. ზედმეტად დიდ n-გრამის მნიშვნელობას შეუძლია გაზარდოს ხმაური, რომელიც გადაფარავს დამოუკიდებელ კონტექსტებს. თუ n-ის ზომა წინადადებაზე დიდია, შეიძლება სულაც არ დააბრუნოს არც ერთი n-გრამი [4,5].

n-ის არჩევა აგრეთვე შესაძლებელია განხილულ იქნეს როგორც კომპრომისი დისპერსიასა და სისტემურ გადახრას შორის. n-ის პატარა მნიშვნელობის შემთხვევაში გამოდის უფრო მარტივი (სუსტი) მოდელი, რაც სისტემური გადახრის გამო იწვევს უფრო მეტ შეცდომას. დიდი ზომის n-ის შემთხვევაში კი გამოდის უფრო რთული მოდელი (უმაღლესი რიგის მოდელი), რაც იწვევს უფრო მეტ შეცდომას დისპერსიის გამო. ამიტომ საჭიროა მოინახოს სწორი ბალანსი მოდელის

მგრძნობელობასა და სპეციფიკას შორის. რაც უფრო მეტი ერთმანეთზე დამოკიდებული სიტყვაა დაშორებული მთავარი სიტყვისგან, მით უფრო რთულია n-გრამის საფუძველზე პროგნოზირებადი მოდელის შექმნა.

მნიშვნელოვანი სიტყვათშეთანხმებები

პრაქტიკაში ზედმეტი n-გრამების დამუშავება, მარტივი აპლიკაციებისთვის, გამოთვლითი რესურსების თვალსაზრისით შესაძლოა ძალიან ძვირი ღირდეს. ამ პრობლემის გადაჭრა შესაძლებელია პირობითი ალბათობის გამოთვლით. მაგალითად, რა არის ალბათობა იმისა, რომ ლეკსემა *"during"* ტექსტში გამოჩნდება (*"the", "fall"*) ლეკსემასთან ერთად. შესაძლებელია ემპირიული ალბათობების პოვნა, (n – 1) გრამის სიხშირის გამოთვლით, რომელსაც n-გრამებიდან წინ უსწრებს პირველი ლეკსემა. ამგვარად შესაძლებელია გავზარდოთ იმ სიტყვების n-გრამების წონები, რომლებიც ხშირად გამოიყენება ერთად.

n-გრამების წონითი ღირებულებების გამოთვლის იდეას, ტექსტის ანალიზში, მივყავართ კიდევ ერთ ინსტრუმენტამდე: **საყურადღებო სიტყვათშეთანხმებამდე** (significant collocations). სიტყვა collocation - ეს არის n-გრამისთვის განზოგადებული სინონიმი (n-ის ზომის მითითების გარეშე) და უბრალოდ ნიშნავს ლეკსემების თანმიმდევრობას, რომელთა ერთდროული გამოჩენის ალბათობა არაა განპირობებული მხოლოდ შემთხვევითი დამთხვევით. პირობითი ალბათობების მეშვეობით შეიძლება შემოწმდეს ჰიპოთეზა არჩეული ფრაზის მნიშვნელოვნობის შესახებ. NLTK ბიბლიოთეკაში არის ორი ინსტრუმენტი სიტყვათშეთანხმებების იდენტიფიცირებისთვის:

CollocatioFinder - იძიებს და აფასებს n-გრამების სიტყვათშეთანხმებას,

NgramAssocMeasures - შეიცავს მეტრიკის კოლექციას სიტყვათშეთანხმების მნიშვნელოვნობის შეფასებისთვის (Wang et al., 2015).

დასკვნა

სწავლება მასწავლებლის გარეშე შესაძლოა არც ისე მარტივი აღმოჩნდეს იმის გამო, რომ არ არსებობს მოდელის ხარისხის შეფასების საიმედო გზა. მიუხედავად ამისა არსებობს მეთოდები, რომლებიც საშუალებას იძლევიან რაოდენობრივად შეფასდეს დოკუმენტების მსგავსება, რომლებსაც შეუძლიათ სწრაფად და ეფექტურად დაამუშაონ დიდი ზომის კორპუსები და წარმოადგინონ საინტერესო და ეფექტური ინფორმაცია.

k-საშუალოს მეთოდი - არის ეფექტური და უნივერსალური კლასტერიზაციის მეთოდი, რომელიც დიდი კორპუსების დამუშავებისათვის კარგად მასშტაბირებადია, განსაკუთრებით თუ კლასტერები არც ისე ბევრია ხოლო გეომეტრია არც ისე რთული. დიდი რაოდენობით კლასტერების შემთხვევაში და ნაკლებად თანაბრად გადანაწილებული მონაცემების კარგ ალტერნატივად შეიძლება აღმოჩნდეს აგლომერაციული კლასტერირება.

კორპუსის დოკუმენტების ეფექტური განზოგადებისთვის, ტეგების გარეშე ხშირად საჭიროა წინასწარი კლასტერირება და კატეგორიების აღწერის მეთოდიც, რასაც აკეთებს თემატური მოდელირება - დირიხლეს ლატენტური ალოკაციის მეთოდით, ლატენტურ-სემანტიკური ანალიზის ან არაუარყოფითი მატრიცული დაყოფით.

ტექსტის სტრუქტურას აქვს დიდი მნიშვნელობა გააზრების თვალსაზრისით. კონტექსტის გამოყენება საკვანძო ფრაზების ამოღების ან საყურადღებო სიტყვათშეთანხმების მეშვეობით საშუალებას იძლევა არსებითად გაუმჯობესდეს მოდელის ხარისხი გრამატიკაზე დაფუძნებით.

გამოყენებული ლიტერატურა

Ali, F., K. Kwak, and Y. Kim. 2016. Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification. Applied Soft Computing 47:235–50. doi:10.1016/j.asoc.2016.06.003.

Bi, J., Y. Liu, Z. Fan, and E. Cambria. 2019. Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. International Journal of Production Research 57(22):1–21.

How Machine Learning Transforms Customer Feedback Analysis URL: https://digitalon.ai/machine-learning-customer-feedback-analysis

Tripathy, A., A. Agrawal, and S. Rath. 2016. Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications 57:117–26. doi:10.1016/j.eswa.2016.03.028.

Wang, F., C. Li, J. Wang, J. Xu, and L. Li. 2015. A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. Journal of Shanghai Jiaotong University (Science) 20:44–50. doi:10.1007/s12204-015-1586-y.

Machine Learning for Processing Customer Feedback Texts in Marketing Tasks

Gulnara Janelidze
Doctor of Engineering Sciences, Associate Professor, GTU

Ia Aptsiauri
Doctor of Engineering in Informatics, Assistant Professor, GTU

Abstract

In modern times, the capabilities of machine learning are rapidly increasing in all areas of human activity, including marketing, without the use and processing of textual information and images. Machine learning has significantly eased the process of working with visual content. Big companies are trying to target and personalize their products. They do this by analyzing people's interests and pulling them in the right direction. It is a proven method that helps organizations to attract a specific audience. In order to maximize the use of investments, a correct orientation to the buyer is required. Without analyzing the audience's wishes in fact, one risks significant losses and customer distrust. The paper presents the use of the clustering algorithm to detect similarities in the text. Problems of context-dependent text analysis based on grammar and feature extraction based on n-grams are presented. The tasks of keyword phrase extraction and total gist detection are discussed. The paper presents the possibilities of using clustering algorithms in marketing, in particular sales tasks, which allows people with similar characteristics to be grouped according to their interests in certain products. At the same time, the clustering algorithm will group products based on customer feedback. As a result, a picture of how much the product is in demand

is obtained, which will be taken into account in order to improve sales. Thus, in sales tasks, processing customer feedback text for subsequent content-analyse is very effective in obtaining product information.

Keywords: detection of similarities in text, selection of n-grams, clustering of text data.
JEL: M3; C45

DOI: 10.52244/c.2024.11.27

References

Ali, F., K. Kwak, and Y. Kim. 2016. Opinion mining based on fuzzy domain ontology and support vector machine: A proposal to automate online review classification. Applied Soft Computing 47:235–50. doi:10.1016/j.asoc.2016.06.003.

Bi, J., Y. Liu, Z. Fan, and E. Cambria. 2019. Modelling customer satisfaction from online reviews using ensemble neural network and effect-based Kano model. International Journal of Production Research 57(22):1–21.

How Machine Learning Transforms Customer Feedback Analysis URL: https://digitalon.ai/machine-learning-customer-feedback-analysis

Tripathy, A., A. Agrawal, and S. Rath. 2016. Classification of sentiment reviews using n-gram machine learning approach. Expert Systems with Applications 57:117–26. doi:10.1016/j.eswa.2016.03.028.

Wang, F., C. Li, J. Wang, J. Xu, and L. Li. 2015. A two-stage feature selection method for text categorization by using category correlation degree and latent semantic indexing. Journal of Shanghai Jiaotong University (Science) 20:44–50. doi:10.1007/s12204-015-1586-y.