საქართველოს ტექნიკური უნივერსიტეტი

ხელნაწერის უფლებით

დავითი გოგშელიძე

საგამომცემლო მარკეტინგის ბიზნეს-პროცესების ავტომატიზაცია სერვის ორიენტირებული არქიტექტურით

სადოქტორო პროგრამა: ინფორმატიკა

შიფრი: 0613

დოქტორის აკადემიური ხარისხის მოსაპოვებლად

წარდგენილი დისერტაციის

ა ვ ტ ო რ ე ფ ე რ ა ტ ი

თბილისი

2022 წელი

სამუშაო შესრულებულია საქართველოს ტექნიკური უნივერსიტეტში

ინფორმატიკისა და მართვის სისტემების ფაკულტეტი

მართვის ავტომატიზებული სისტემების (პროგრამული ინჟინერიის) დეპარტამენტი

ხელმძღვანელი: პროფესორი გია სურგულაძე

რეცენზენტები: პროფესორი ირინა ხომერიკი
პროფესორი დიმიტრი მასხარაშვილი

დაცვა შედგება 2022 წლის ”28” ივლისს, 14:00 საათზე

საქართველოს ტექნიკური უნივერსიტეტის ინფორმატიკისა და მართვის სისტემების ფაკულტეტის სადისერტაციო ნაშრომის დაცვის კოლეგიის სხდომაზე,
კორპუსი ----------, აუდიტორია ----------

მისამართი: 0160, თბილისი, კოსტავას 77.

დისერტაციის გაცნობა შეიძლება სტუ-ის ბიბლიოთეკაში,

ხოლო ავტორეფერატისა - ფაკულტეტის ვებგვერდზე

ფაკულტეტის სწავლული მდივანი ----------

თემის აქტუალურობა

წარმოდგენილი სადისერტაციო ნაშრომის საკვლევი თემა, ეხება საგამომცემლო მარკეტინგისა და ტექსტური ინფორმაციის შექმნის ბიზნეს პროცესების გაუმჯობესებას. ტექნოლოგიური ევოლუციის პარალელურად, ტექსტური ინფორმაციის ელექტრონული გამოქვეყნება და კომერცია საგრძნობლად იზრდება და ვითარდება. არსებობს უამრავი ელექტრონული ბიბლიოთეკა და ონლაინ მაღაზია, სადაც მილიონობით ნაშრომსა თუ წიგნს შეხვდებით. ამასთან ერთად, არსებობს განვითარებული მრავალფუნქციური ტექსტური რედაქტორები, რომლებიც ხელს უწყობენ ტექსტური ინფორმაციის შექმნას. თუმცა, ასეთი ტექნოლოგიური და ავტომატიზებული სისტემების განვითარების ფონზეც კი შეგვიძლია დარწმუნებით ვთქვათ, მისი განვითარების გზა ჯერ კიდევ წინაა. დღეს არსებული ტექსტური რედაქტორები, მორგებულია ყველა ტიპის ინფორმაციის შექმნაზე ერთდროულად, მწერალს კი ნაკლები შანსი აქვს მიჰყვეს მისი წერის მეთოდოლოგიას. პროექტის ერთ-ერთი ინოვაციურობა სწორედ ამ პრობლემის მოგვარებაში მდგომარეობს. დღესდღეობით, არსებობს მრავალი დოკუმენტური ფაილების შექმნის პროგრამა, თუმცა არ არსებობს სპეციალურად სამეცნიერო ნაშრომების შექმნისათვის განკუთვნილი პროგრამული პროდუქტი, რომელიც სხვადასხვა სამეცნიერო მიმართულების მქონე მეცნიერთათვის, სპეციალურად მათზე მორგებული ფუნქციონალისა და ინტერფეისის შეთავაზებით იქნება ხელშემწყობი. სადოქტორო პროექტის მიზანია, რომ მანქანური სწავლებების ალგორითმების გაშვების გზით შეიქმნას ისეთი მოდელები, რომელთა პრაქტიკული გამოყენებითაც, ტექსტური ინფორმაციის შექმნისა და საგამომცემლო მარკეტინგის ბიზნეს პროცესების გამარტივება იქნება შესაძლებელი.

მეცნიერული სიახლე

სადოქტორო კვლევის ერთ-ერთი მთავარი მეცნიერული სიახლე, თემის მოდელირების ინოვაციური მოდელების მიღებაა. სადოქტორო კვლევისთვის არსებული მონაცემთა ბაზის შესწავლისას შედარებული იქნა თემის მოდელირების ორი უმნიშვნელოვანესი მეთოდი: დირიხლეს ლატენტური განაწილება და არაუარყოფითი მატრიცის ფაქტორიზაცია. ამ მეთოდების მიერ საბოლოო შედეგის მიღებამდე, ჩვენს მიერ მოხდა მათი პარამეტრების ოპტიმიზაცია, მათი არაერთგზის გაშვების გზით. თემის მოდელირება მოხდა სამეცნიერო სტატიების რეზიუმეებისა და მათი სრული ტექსტებისათვის. საბოლოოდ, კი შედეგად მიღებულია, ორი ინოვაციური მოდელი, რომლის საშუალებითაც, მოხდა არსებული სამეცნიერო ნაშრომების რეკატეგორიზაცია. მათი ცალკეული ისევე როგორც სინქრონული გამოყენება, საშუალებას იძევა რომელიმე კატეგორიის სამეცნიერო მიმართულებით და ამ მიმართულებაში კონკრეტული თემით დაინტერესებული პირებისთვის, მოხდეს მათთვის საინტერესო ნაშრომების შევთავაზება. ხსენებული მეთოდების გამოყენებით მიღებული მოდელები ინოვაციურია და შედეგი წარმოადგენს სამეცნიერო სიახლეს.

აღსანიშნავია მანქანური სწავლების მეორე მეთოდის, ტექსტის კლასიფიკაციის გამოყენებაც. სამეცნიერო ნაშრომების კლასიფიკაციის მოდელების შესაქმნელად, გამოვიყენებული იქნა მანქანური სწავლების ორი ყველაზე განვითარებული მეთოდი: ხაზოვანი მხარდაჭერის ვექტორული კლასიფიკატორი (SVC) და მულტინომინალური მიამიტი ბეიესის ალგორითმი. ამ მეთოდების გამოყენების შედეგად, მიღებული იქნა მოდელი, რომლის მეშვეობითაც, რამდენიმე წინადადების შეყვანის შემდეგ, შესაძლებელია მომენტალურად დადგინდეს ის, თუ რა სფეროს ეხება ან უბრალოდ რა თემაზეა აღნიშნული ტექსტი. ამის მიხედვით, კი შესაძლებელი იქნება, რომ სამეცნიერო ნაშრომის მწერალ პირთათვის, მათ სფეროზე მორგებული ფუნქციონალისა და მომხმარებლის ინტერფეისის შევთავაზება მოხდეს.

ნაშრომის სამეცნიერო სიახლე გახლავთ გამოკითხვებისა და სტატიებში გამოყენებული ფუნქციონალის პროცენტული განაწილების კვლევის შედეგებიც. აღნიშნული კვლევიდან გამომდინარე, მივიღებულ იქნა ინფორმაცია იმის შესახებ, თუ რა ფუნქციონალი და ინტერფეისია საჭირო, ტექსტურ ინფორმაციასთან მომუშავე პირთა მუშაობის გამოცდილების გასამარტივებლად. ამასთან ერთად, შესწავლილი იქნა ბიზნეს პროცესები, რომელიც სამეცნიერო ნაშრომებთან მუშაობის დროს მიმდინარეობს. ამ ცონდის საფუძველზე კი ჩამოყალიბებულ იქნა მათი ავტომატიზაციის მოდელები. სამეცნიერო ნაშროებში არსებული სხვადასხვა ფორმების პროცენტული განაწილების კვლევის შემდეგ, მიღებულია ცოდნა იმასთან დაკავშირებით, ყველზე მეტად რა ტიპის ფუნქციონალია გამოყენებული სხვადასხვა კატეგორიის სამეცნიერო ნაშრომში. ამ მონაცემების საფუძველზე კი ჩამოყალიბებულ იქნა პროტოტიპი, რომლის მეშვეობითაც თვალსაჩინოდ ჩანს ის, თუ რას ნიშნავს, სხვადასხვა ტიპის სამეცნიერო მიმართულებისთვის მორგებული მომხმარებლის ინტერფეისი და ფუნქციონალი.

სამუშაოს მიზანი

კვლევის მთავარ მიზანია სხვადასხვა სფეროს მწერლებისთვის საგამომცემლო მარკეტინგისა და ტექსტური ინფორმაციის შექმნის პროცესში ხელშეწყობაა. ეს კი მათი მუშაობის სტილისა და მეთოდოლოგიის გათვალისწინებით უნდა მოხდეს. კვლევის ერთ-ერთი მიზანია, რომ პროექტში მოდელირებულმა სისტემამ მუშაობის პროცესი გაუმარტივოს მთარგმნელებსაც.

აგრეთვე, სადოქტორო პროექტის მიზანია შეიქმნას ჩამოყალიბებული, დოკუმენტირებული ინფორმაცია, კვლევის შედეგად მიღებული სამეცნიერო სიახლის პოტენციური გამომყენებელი პროგრამის არქიტექტურის, მისი დიზაინის, კლასებისა და მოდელების შესახებ. გაანალიზებულ იქნას ის, თუ რა

კომპიუტერული რესურსები, პროგრამული ენები, მონაცემთა ბაზები, თუ სხვა ტექნიკური საშუალებებია საჭირო პროექტის შესრულებისთვის.

კვლევის ობიექტი და მეთოდები

სადოქტორო კვლკვლევის ობიექტია საგამომცემლო სფერო და მასში არსებული ბიზნეს პროცესების გაუმჯობესება. ამისათვის გამოყენებული იქნა კვლევის სამი მეთოდი:

1. მანქანური სწავლების ალგორითმები.
2. კონტექსტუალური გამოკითხვები.
3. სხვადასხვა ტიპის სამეცნიერო სტატიების განხილვა.

მანქანური სწავლების ეტაპზე, გამოყენებული იქნა ბუნებრივი ენის დამუშავების ორი მეთოდი, თემის მოდელირება და ტექსტის კლასიფიკაცია. თემის მოდელირებისთვის გამოყენებულ იქნა დირიხლეს ლატენტური განაწილება და არაუარყოფითი მატრიცის ფაქტორიზაციის ალგორითმები. ტექსტის კლასიფიკაციის მთავარ შემსწავლელ ალგორითმებად კი ხაზოვანი მხარდაჭერის ვექტორული კლასიფიკატორი და მულტინომინალური მიამიტი ბეიესის ალგორითმი გამოვიყენეთ.

კონტექსტუალური გამოკითხვებისას დასმული შეკითხვები იყო შედგენილი ისე, რომ მიღებულ იქნა ინფორმაცია, ტექსტური ინფორმაციის შექმნისა და მასთან მუშაობის სხვა ბიზნეს პროცესების დროს არსებულ პრობლემებზე. ჩვენს მიერ გაანალიზდა ტექსტური ინფორმაციის შექმნისათვის საჭირო ისეთი ფუნქციონალის საჭიროება, რომელიც დღევანდელ ტექსტურ რედაქტორებს არ გააჩნიათ. აგრეთვე, მიღებული იქნა ინფორმაცია, სფეროში არსებული ბიზნეს პროცესებზე და ამ პროცესების ავტომატიზაციის საჭიროებაზე. ამის მიხედვით ჩამოყალიბებული იქნა ამ პროცესების ავტომატიზაციის გზები.

ამასთან ერთად, განხილული იქნა დღესდღეობით ბაზარზე არსებული მიდგომები, ტექსტური ინფორმაციის შექმნისა და მისი გამოყენებისთვის. აგრეთვე ჩატარდა ანალიზი იმის შესახებ თუ რა პროგრამულ, აპარატურული

რესურია საჭირო კვლევების მეორე და მესამე ეტაპზე მიღებული ინფორმაციის შესასრულებლად.

კვლევის ძირითადი შედეგები და შედეგების გამოყენების სფერო

კვლევის შედეგად მიღებული იქნა თემის მოდელირების ორი მოდელი, სამეცნიერო ნაშრომთა რეზიუმეებისა და მისი სრული ტექსტებისთვის. მანქანამ შეისწავლა ორ მილიონამდე სამეცნიერო სტატიის რეზიუმე და ასევე ოთხასი ათასამდე სამეცნიერო სტატიის სრული ტექსტი. სწავლებისთვის საჭირო მონაცემთა ოპტიმიზაციისა და სხვადასხვა მეთოდების გამოყენების შემდეგ, მიღებული იქნა ორი ინოვაციური ზემოთ ხსენებული მოდელი, რომელთა სინქრონულად გამოყენებაც ძალიან კარგად აისახება სამეცნიერო საგამომცემლო სფეროზე. ამ მოდელების გამოყენებით, შესაძლებელია ადამიანის სამეცნიერო ინტერესის გამოვლენა და მისთვის საინტერესო ნაშრომების შეთავაზება. იმდენად, რამდენადაც რეზიუმეები იძლევიან შედარებით ზოგად ინფორმაციას სამეცნიერო სფეროს შესახებ, სტატიების სრული ტექსტი იძლევა ინფორმაციას, ამ სფეროში ჩატარებული კვლევის მეთოდოლოგიებთან დაკავშირებით. მანქანური სწავლებების მოდელების გამოყენებით, შესაძებელი ხდება, რომ ადამიანებისთვის შევთავაზებული იქნას მათთვის საინტერესო სფეროს ნაშრომები, რომლებიც მათთვის საინტერესო მეთოდოლოგიითაა შესრულებული.

კვლევის კიდევ ერთი შედეგია, ტექსტის კლასიფიკაციის მეთოდით მიღებული მოდელი, რომლის სიზუსტის მაჩვენებელიც ორასამდე, ხშირად ერთმანეთთან ახლოს მყოფი, კატეგორიისათვის 60%-ია. აღნიშნული მაჩვენებელი სადოქტორო კვლევისთვის ძალიან მაღალი და შედეგისთვის დამაკმაყოფილებელია. მიღებული მოდელის გამოყენებით, როდესაც მომხმარებელი ტექსტურ რედაქტორში მუშაობას დაიწყებს, რედაქტორი პირველი წინადადებების მიღების შემდეგ, მომენტალურად შეძლებს

გამოიცნოს ის, თუ რა თემაზე იქმნება ნაშრომი და მის შემქმნელს მისთვის საჭირო სამუშაო გარემოს შეუქმნის. ის მას სფეროში ხშირად გამოყენებულ ფუნქციონალს შესთავაზებს და მომხმარებლის ინტერფეისს მისთვის სასარგებლოდ განალაგებს. იმის გასაგებად, თუ რა ფუნქციონალი და ინტერფეისია ამ მომხმარებელთათვის საჭირო და მნიშვნელოვანი, კვლევების შემდეგ ეტაპებშია გამოვლენილი.

კვლევის შემდეგ ეტაპზე გამოვლინდა ინფორმაცია, საგამომცემლი სფეროში არსებული ბიზნეს პროცესების ავტომატიზაციის საჭიროებების შესახებ. შედეგად კი ჩამოყტალიბებული იქნა ამ პროცესების ავტომატიზაციის მოდელები. აგრეთვე უნდა აღინიშნოს, რომ მიღებულ იქნა ინფორმაცია იმ ფუნქციონალის საჭიროებების შესახებ, რომელიც ნაშრომის შექმნისთვის მნიშველოვანია, მაგრამ დღევანდელ ტექსტურ რედაქტორებს არ გააჩია.

გამოვიკვლეული იქნა სხვადასხვა კატეგორიის მქონე სამეცნიერო ნაშრომებში გამოყენებული ფუნქციებისა და ფორმების რაოდენობრივი მაჩვენებლები. მიღებული მონაცემები კი კვლევის ბოლო ეტაპზე წარმოდგენილ პროტოტიპში იქნა გამოყენებული, რომელიც სხვადასხვა კატეგორიის სამეცნიერო მიმართულების მომხმარებელს მწერლებს, მათზე მორგებულ მომხმარებლის ინტერფეისს სთავაზობს.

ცნობები დისერტაციის მოცულობისა და სტრუქტურის შესახებ

დისერტაცია შედგება 124 გვერდისაგან. შეიცავს რეზიუმეს ქართულ და ინგლისურ ენებზე. შედგება შესავალისგან, 6 თავის: 1. თემის მიზანი, 2. სადოქტორო ნაშრომისათვის საჭირო ტექნიკური ინსტრუმენტები, 3. საგამომცემლო სფეროში არსებული პროდუქტების მიმოხილვა, 4. მანქანური სწავლება - კვლევა და შედეგები, 5. ნაშრომის შექმნისათვის საჭირო ფუნქციონალისა და ინტერფეისი კვლევა და მისი შედეგები, 6. პროტოტიპი და დასკვნისაგან. ნაშრომი შეიცავს 34 ნახაზსა და 3 ცხრილის. შეიცავს

გამოყენებული ლიტერატურის ნუსხას, რომელშიც წარმოდგენილია 44 წყარო.

დისერტაციის ძირითადი შედეგები თავების მიხედვით

შესავალი

დისერტაციის შესავალში საუბარია თემის აქტუალობაზე. განხილულია ძირითადი იდეა და მისი გამოყენების შედეგად მიღებული პოტენციური სარგებელი. კვლევის შედეგად მიღებული პოტენციური სარგებლის შესახებ აღნიშნულია, რომ ნაშრომში გამოყენებული მეცნიერული კვლევების საფუძველზე, შესაძლებელი გახდება, სხვადასხვა სფეროს მეცნიერთათვის, წერის პროცესში, სპეციალურად მათ სფეროზე მორგებული ინტერფეისია და ფუნქციონალის შეთავაზება. აგრეთვე აღნიშნულია რომ კვლევების შედეგად მიღებული ინფორმაციის საფუძველზე, შექმნილია სხვადასხვა ტიპის ნაშრომზე მუშაობის დროს არსებული ბიზნეს პროცესების ავტომატიზაციის მოდელები. ამასთან, აღნიშულია, რომ მანქანური სწავლების შედეგად მიღებული მოდელების მეშვეობით, გაუმჯობესდება სამეცნიერო ლიტერატურის საგამომცემლო მარკეტინგის მიმართულება. შესავალში აგრეთვე აღნიშნულია საგამომცემლო სფეროში მთარგმნელთა როლის შესახებ.

1. თემის მიზანი

პირველი თავში, გადმოცემულია სადოქტორო ნაშრომის მიზანი, მეთოდოლოგია, სამეცნიერო სიახლე და პრაქტიკული ღირებულება, ინტერდისციპლინურობა და არეალი.

აღნიშნულ თავში გადმოცემულია, რომ კვლევა დაყოფილია სამ ნაწილად. მისი პირველი ეტაპი ტექნიკურია და ეხება იმის განსაზღვრას, თუ რა ტექნიკური საშუალებებია საჭირო კვლევის ჩასატარებლად. მეორე ეტაპზე, ჩამოყალიბებული იქნა ინოვაციური მოდელები მანქანური სწავლების მეთოდების გამოყენებით. აგრეთვე, ჩამოყალიბებულ იქნა ბიზნეს მოთხოვნა იმისა თუ რა ფუნქციონალი და შესაძლებლობებია საჭირო სადოქტორო

პროექტში დასახული მიზნების შესასრულებლად. მესამე ეტაპზე კი პირველ პუნქტში შერჩეული ტექნიკური საშუალებებითა და მეორე პუნქტში მიღებული კვლევის შედეგებით, წარმოდგენილი იქნა პროტოტიპი, რომელიც ასახავს ტექსტური ინფორმაციის შექმნის ფუნქციონალს, მისი შესრულების გზებსა და მის შესაძლებლობებს.

2. პროექტის ტექნიკური ინსტრუმენტები

მეორე თავში, საუბარია პროექტის ტექნიკური ინსტრუმენტების შესახებ. განხილულია ბუნებრივი ენის დამუშავება, თემის მოდელირება, ტექსტის კლასიფიკაცია და სხვა მანქანური სწავლებების მუშაობის პრინციპები. საუბარია პროექტის ისეთი ტექნიკური დეტალების შესახებ, როგორიცაა სერვერზე მონაცემთა განთავსება, პროგრამული ენები და ბიბლიოთეკები, პროგრესული ვებ აპლიკაცია და სხვა.

მანქანური სწავლება, თავის მხრივ ხელოვნური ინტელექტის ქვე-კატეგორიაა. ის არის მონაცემთა მეცნიერების ერთ-ერთი უმნიშვნელოვანესი ნაწილი, რომელიც კომპიუტერს რეალობის პროგნოზირების საშუალებას აძლევს. მისი ძირითადი მიზანი გახლავთ ის, რომ ტექნოლოგიურ მანქანას შეეძლოს მიბაძოს ადამიანის ქცევებსა და აღქმას, სხვადასხვა მიმართულებით, მათ შორის არსებული რეალობის გააზრებითა და გაანალიზებით.

მანქანური სწავლება სხვადასხვა პროგრამებს ეხმარება პირდაპირი დაპროგრამების გარეშე, მოახდინონ ამა თუ იმ რეალობის პროგნოზირება. ამას კი ისინი სხვადასხვა ალგორითმებისა და მათი გამოყენების შედეგად მიღებული მოდელების, მიერ შესწავლილი ისტორიული მონაცემების გამოყენებით ახდენენ.

მოცემულ მეცნიერებაში, სამი ძირითადი მეთოდი არსებობს. ესენია: კონტროლირებადი სწავლება, არანოკტროლირებადი სწავლება და ნახევრად კონტროლირებადი სწავლება. კონტროლირებადი მანქანური სწავლება

განისაზღვრება ისეთი მონაცემთა ნაკრებით, რომლებიც რაიმე გზით არიან კლასიფიცირებულნი და გააჩნიათ ინფორმაცია ამ კლასების შესახებ. ნაშრომში ამ ინფორმაციის გამოყენება ხდება სხვადასხვა პროგრამული მიდგომების მეშვეობით ისე, რომ საბოლოოდ შესაძლებელია ახალი შეყვანილი მონაცემების კლასიფიკაცია ან პროგნოზი. არაკონტროლირებადი მანქანური სწავლება იყენებს სხვადასხვა მათემატიკურ ფუნქციებსა და პროგრამულ ალგორითმებს იმისთვის, რომ მოახდინოს ისეთი მონაცემთა კატეგორიზაცია, რომელსაც ჯერ არ გააჩნიათ რაიმე კატეგორია. მაგალითად, სადოქტორო კვლევის შემთხვევაში, ალგორითმს მივეცით მილიონობით სამეცნიერო სტატიის რეზიუმეს ტექსტი და მისი საშუალებით მოხდა ამ ნაშრომების ახალ კატეგორიებში განთავსება. შესაბამისად, ეს ალგორითმები ახდენენ მონაცემთა ავტომატურ დაჯგუფებას ადამიანის ჩართულობის გარეშე. ნახევრად-კონტროლირებადი მანქანური სწავლება, კი არის მიდგომა კონტროლირებად და არა კონტროლირებად მანქანურ სწავლებებს შორის, რომელიც თავის მხრივ მოიცავს ორივე მათგანს.

იქამდე, სანამ თემის მოდელირებისა და კლასიფიკაციის ალგორითმები იქნა შემუშავებული, მოხდა არსებული სამეცნიერო ნაშრომებისა და რეზიუმეების გასუფთავება ისეთი ზედმეტი ტექსტებისგან, რომლებმაც შეიძლება ხელი შეუშალოს ორივე ტიპის მანქანურ სწავლებას. ასეთი ტექსტები შეიძლება იყოს, არტიკლები, კავშირები, რიცხვები და სხვა ისეთი სიტყვები ან ფრაზები, რომლებიც კონკრეტულად ამ სადოქტორო კვლევისთვისაა ხელის შემშლელი ფაქტორი. ამ პრობლემების აღმოსაფხვრელად, მანქანურ სწავლებაში არსებობს მიდგომა - ბუნებრივი ენის დამუშავება (NLP).

თემის მოდელირება არის მანქანური სწავლების არა-კონტროლირებადი მეთოდი, რომლის მეშვეობითაც ხდება სხვადასხვა ტიპის ტექსტური ინფორმაციისგან შემდგარი დოკუმენტების კატეგორიზაცია მასში არსებული ტექსტის, კერძოდ კი გამოყენებული სიტყვებისა და ფრაზების მიხედვით. მეთოდი ავტომატურად აანალიზებს ტექსტურ მონაცემებს და წინასწარ

მითითებული კლასტერების რაოდენობის მიხედვით ანაწილებს ამ კლასტერებში. საბოლოო შედეგად კი მიიღება წინასწარ მითითებული რაოდენობის კლასტერი/კატეგორია და მასში მოხვედრილი ტექსტურ მონაცემები.

თემის მოდელირება სასარგებლოა იმაში, რომ სხვადასხვა ტიპის დოკუმენტები საერთო კლასტერის ქვეშ დაჯგუფდეს, მასში არსებული სიტყვებისა და ფრაზების მიხედვით. თემის მოდელირების ერთ-ერთი ყველაზე პოპულარული მიდგომა, დირიხლეს ლატენტური განაწილება იქნა გამოყენებული. დირიხლეს ლატენტური განაწილება (LDA) არის თემის მოდელირების, არაუარყოფითი მატრიცის ფაქტორიზაციის მეთოდთან ერთად, ყველაზე განვითარებული ტექნიკა, დოკუმენტთა კრებულიდან სხვადასხვა თემების ამოსაღებად. ამოსაღები თემები, რომლებიც ალგორითმის გაშვების შემდეგ უნდა გენერირდეს, ჯერ კიდევ უცნობია. სწორედ ამიტომ გამოიყენება სიტყვა ლატენტური, რაც გულისხმობს რაიმეს, რაც უკვე არსებობს, მაგრამ ჯერ არ არის განვითარებული.

არაუარყოფითი მატრიცის ფაქტორიზაცია, თემის მოდელირების ერთ-ერთი ალგორითმი და შესაბამისად, მანქანური სწავლების არაკონტროლირებადი მიდგომაა. ის ერთდროულად ახდენს როგორც განზომილებათა შემცირებას, ასევე მის კლასტერიზაციას. მოცემული ალგორითმის გამოყენება, შესაძლებელია TF-IDF (ტერმინი სიხშირე-შებრუნებული დოკუმენტის სიხშირე) ვექტორიზერთან ერთად. TF-IDF გახლავთ სტატისტიკური საშუალება იმისა, რომ მოხდეს გაზომვა თუ რამდენად შეესაბამება ესა თუ ის სიტყვა რომელიმე კონკრეტულ დოკუმენტს, სხვადასხვა დოკუმენტთა სიმრავლეში. TF-IDF მიიღწევა ორი მნიშვნელობის გამრავლებით, ესენია: ტერმინთა სიხშირე დოკუმენტში - გულისხმობს იმას, თუ რამდენადაა სიტყვა გამოყენებული კონკრეტულ დოკუმენტში და დოკუმენტის ინვერსიული სიხშირე - გულისხმობს იმას თუ რამდენად ხშირადაა ესა თუ ის სიტყვა გამოყენებული მთლიან დოკუმენტთა სიმრავლეში.

კიდევ ერთი მნიშვნელოვანი მანქანური სწავლების მეთოდი, რომელსაც სადოქტორო კვლევისათვის იქნა გამოყენებული ტექსტის კლასიფიკაციაა. კლასიფიკაცია, მანქანური სწავლების, ერთ-ერთი უმთავრესი და ყველაზე პოპულარული ტექნიკაა. შესაბამისად, ტექსტის კლასიფიკაციაც ფუნდამენტური მიდგომაა ბუნებრივი ენის დამუშავების სფეროში და შექმნილი იქნა მოდელი იმისთვის, რომ გამოყენებული იქნას ისეთი დავალებების შესასრულებლად, როგორიცაა ახალი ნაშრომისთვის კატეგორიის ავტომატური მინიჭება და ნაშრომზე მუშაობის დაწყებისას მისი კატეგორიის გამოცნობა. მოცემული მეთოდი კონტროლირებადია და ტრენინგისათვის საჭიროებს მარკირებულ მონაცემებს (მონაცემებს უკვე არსებული კატეგორიით).

მანქანურ სწავლებაში, ტექსტის კლასიფიკაციის სხვადასხვა მეთოდი არსებობს. მოცემულ სადოქტორო კვლევაში რამოდენიმე ყველაზე პოპულარული მეთოდი იქნა გამოყენებული, ხოლო შედეგები ერთმანეთს შედარებული, და საბოლოოდ შეირჩა სასურველი მეთოდი. სადოქტორო პროექტში გამოყენებული პოპულარული ალგორითმებია: მიამიტი ბეიესის ალგორითმთა ოჯახი და დამხმარე ვექტორული მანქანები (SVM).

მეორე თავში მანქანური სწავლების მეთოდებთან ერთად, საუბარია სერვერული მონაცემების განთავსებაზე. უნდა აღინიშნოს, რომ სერვერზე წვდომის სისწრაფე პირდაპირ პროპორციულია, მოთხოვნის გამშვების მასთან გეოგრაფიულად ახლოს მყოფობასთან. თუ სერვერი ტერიტორიულად მდებარეობს ჩრდილოეთ ამერიკაში და მასზე წვდომა ავსტრალიიდან ხდება, ინფორმაციის გაცვლის სისწრაფე მანძილის შესაბამისად დაბალია. დღევანდელი მსოფლიოს თანამედროვე მიღწევები, მოცემული პრობლემის გადაჭრას ღრუბლოვანი სერვისების საშუალებით იძლევიან (მაგ. AWS). დასაწყისისთვის საჭიროა გამოყენებულ იქნას ყველაზე კომფორტული ღრუბლოვანი სერვისი. სადოქტორო პროექტის შემთხვევაში ეს სერვისი

საშუალებას უნდა იძლეოდეს, თავისუფლად მოხდეს მონაცემთა ბაზებისა და სერვერის აპლიკაციის უსაფრთხოდ განთავსება, უნდა იყოს დამხმარე ავტომატური ტესტირებსა და CI/CD მენეჯმენტში. დღესდღეობით, მსოფლიოში ერთ-ერთი ყველაზე გავრცელებული და სანდო ღრუბლოვანი სერვისია AWS, იგი გვთავაზობს ბევრ კვლევისათვის მნიშვნელოვან და საინტერესო ღრუბლოვან სერვისს, როგორებიცაა AWS S3 (Simple Storage Service), EC2 (Elastic Compute Cloud) და სხვა. რაც შეეხება მონაცემთა ბაზებს, სადოქტორო პროექტის სრულყოფილი მუშაობისთვის საჭიროა სამი სახის მონაცემთა ბაზები. რელაციური MySQL - მონაცემთა ბაზები. Google Firebase რეალურ დროში განახლებადი მონაცემთა ბაზები და, არარელაციური იგივე noSql მონაცემთა ბაზები. მეორე თავში, ასევე ვსაუბრობთ გამოყენებული პროგრამული ენებისა და ბიბლიოთეკების შესახებ. მანქანური სწავლების ალგორითმების გასაშვებად, ზოგადად პროგრამირების ენა python-ი და მასში არსებული ბიბლიოთეკები იქნა გამოყენებული, ხოლო რაც შეეხება დამატებით ბიბლიოთეკებს, ამ შემთხვევაში მოკვლეული და გამოყენებული იქნა ნაშრომის მოთხოვნების შესასრულებლად საჭირო, მსოფლიოში საუკეთესო ღია კოდის მქონე ბიბლიოთეკები. პირველ რიგში აღსანიშნავია „sklearn“ ბიბლიოთეკა. მანქანური სწავლებისთვის განკუთვნილი მისი უამრავი ფუნქციონალის საშუალებით, სამეცნიერო კვლევის პროცესი ბევრად უფრო სწაფი აღმოჩნდა ვიდრე მის გარეშე იქნებოდა. ბიბლიოთეკა გამოყენებული იქნა, როგორც მონაცემთა წინასწარი მომზადებისთვის, აგრეთვე მანქანური სწავლების ალგორითმების გასაშვებად და მისი ტესტირებისთვის. აგრეთვე, აღსანიშნავია, რომ მიღებული მოდელების შესანახად გამოვიყენენული იქნა ბიბლიოთეკა „pickle“, ხოლო შედეგების ვიზუალიზაციისთვის ბიბლიოთეკა „pyLDAvis“ და გამოცდილი „matplotlib“-ი.

ამავე თავში შერჩეული იქნა პროგრამული ენები და ბიბლიოთეკები, ვებ აპლიკაციის - კლიენტის და სერვერის აპლიკაციისთვის. არჩეული

ტექნოლოგიებია შესაბამისად, Angular და Nodej. მეორე თავში ასევე საუბარია პროგრესული ვებ აპლიკაციებისა და მისი უპირატესობების შესახებ.

3. არსებული პროდუქტების მიმოხილვა

მესამე თავში საუბარია არსებულ პროდუქტებზე, ტექსტური ინფორმაციის შექმნის, მისი წაკითხვისა და გამოქვეყნებისთვის. საუბარია იმაზე, რომ დღესდღეობით არსებული ტექსტური რედაქტორები, განსაკუთრებით კი Microsoft Word-ი საკმაოდ განვითარებულია და მასზე მიჩვევადობის მაღალი მაჩვენებლით გამოირჩევა. აქედან გამომდინარე, მიღებულია დასკვნა, რომ საჭიროა შესწავლილ იქნას მისი თავისებურებები, რომ მომხმარებელს ახალი ინტერფეისზე მუშაობის დროს მიჩვევადობის გამო არ შეექმნას მუშაობის პრობლემა. თუმცა, აღვნიშნავთ, რომ კვლევის დაწყებამდე, იდეის ვალიდაციის მიზნით ჩატარებული კვლევების მიხედვით, შეიძება ითქვას, რომ დღესდღეობით, ბაზარზე არ არსებობს ისეთი სისტემა, რომელიც სხვადასხვა კატეგორიის სამეცნიერო ნაშრომებისთვისაა შექმნილი და ამ კატეგორიაზე მომუშავე მეცნიერებს, მათზე მორგებულ ინტერფეისსა და ფუნქციონალს სთავაზობს. ამასთან ერთად, დაეხმარება მას სტატიის შემოწმებისა და მისი დასრულების ავტომატიზაციაში, თანაავტორებთან ურთიერთობაში და ჩასწორებების ისტორიის შენახვაში. არ არსებობს სისტემა, რომელიც სამეცნიერო ნაშრომის დაწყებისას პირველი წინადადებების შეყვანის შემდეგ, მანქანური სწავლების მეთოდებით მიღებული მოდელის საშუალებით გამოიცნობს ნაშრომის თემას და მწერალს სპეციალურად მასზე მორგებულ იტერფეისს შესთავაზებს.

მოცემულ თავში აგრეთვე საუბარია არსებული წიგნების გამოცემისა და მისი წაკითხვის შესაძლებლობებზე.

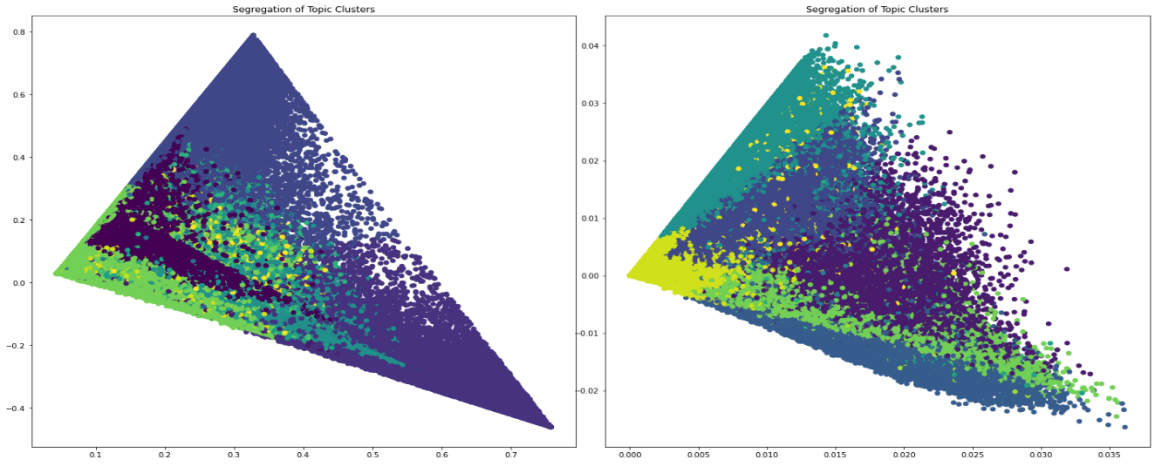
4. მანქანური სწავლება - კვლევა და შედეგები

მეოთხე თავში, განხილულია მანქანური სწავლების კვლევები და ამ კვლევათა შედეგები. სადოქტორო პროექტის სამეცნიერო ღირებულებისა და სამეცნიერო სიახლის უმნიშვნელოვანესი ნაწილი, მანქანური სწავლების მეთოდებისა და ალგორითმების გამოყენებით ინოვაციური მოდელების შექმნაა. პირველ რიგში, ნებისმიერი მანქანური სწავლების ალგორითმისთვის აუცილებელია სასწავლო მონაცემთა ბაზა (Dataset). რაც უფრო დიდია მონაცემთა ბაზა, მით უფრო უკეთესია შედეგიც. ნაშრომისთვის მოკვლეულ იქნა Arxiv.org-ის მონაცემთა ბაზა, რომელიც ორ მილიონამდე სამეცნიერო ნაშრომს მოიცავს. ამასთან დამატებული იქნა Pubmed-ის მონაცემთა ბაზა, რომელიც 130000-მდე სამეცნიერო სტატიისგან შედგება.

როგორც უკვე აღვნიშნეთ, კვლევაში გამოყენებულია ორი LDA და NMF მეთოდი. მაგრამ იქამდე, სანამ თემის მოდელირების ალგორითმს იქნება გაშვებული, საჭიროა მონაცემთა წინასწარი დამუშავება. გარკვეული შეზღუდვების გამო, LDA და NMF, მონაცემთა წინასწარი დამუშავება განსხვავებული მეთოდებით ჩატარდა. იმის გამო, რომ LDA დამოკიდებულია სიტყვათა რაოდენობის ალბათობათა განაწილებაზე, არ შეგვიძლია TF-IDF-ს ვექტორიზერის გამოყენება, განსხვავებით NMF მოდელისგან, რომელიც კოეფიციენტების საშუალებით მუშაობს. LDA-ში გამოყენებული იქნა დათვლის შედარებით მარტივი CountVectorizer ვექტორიზერი. “fit_transform” მეთოდის გაშვებამდე, რომელიც საბოლოოდ გარდაქმნის დოკუმენტებს სიტყვათა ვექტორებად, ორივე ტიპის ვექტორიზერის გენერირების დროს გამოყენებულია მასში შემავალი სამი პარამეტრი. ესენია: “max_df” – პარამეტრის გამოყენებით, მონაცემთა სიმრავლეებში ხდება ფილტრაცია (იგნორი) ისეთ სიტყვებსა, რომლებიც მითითებული რიცხვის პროცენტულ რაოდენობაშია გამოყენებული. სადოქტორო კვლევის შემთხვევაში, ეს რიცხვი 0.5-ია. შესაბამისად, ყველა ის სიტყვა, რომელიც სამეცნიერო ნაშრომთა 50%-შია გამოყენებული, თემის მოდელირების დროს იგნორირებული იქნა. “min_df” – წინა პარამეტრის

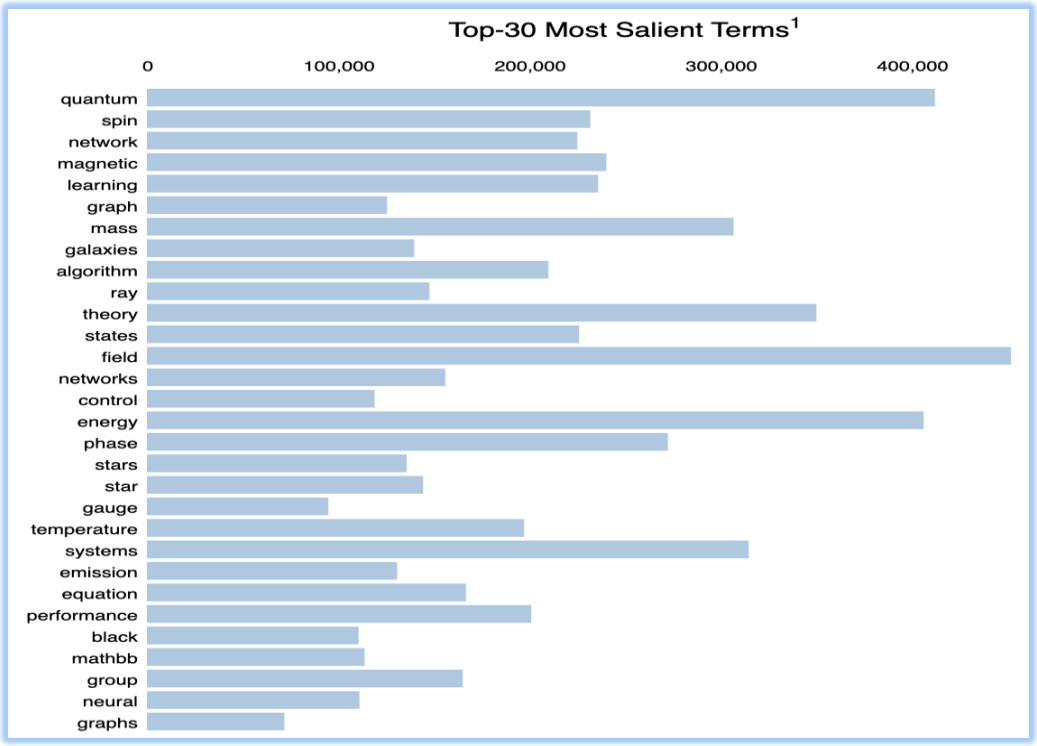
საპირისპიროდ, მოცემული პარამეტრის მნიშვნელობით იფილტრება სიტყვები, რომლებიც მინიმუმ 2 სტატიაშია გამოყენებული. “stop_words” – კი მეთოდისათვის დასაიგნორებელ სიტყვათა სიის გადასაცემად არის გამოყენებული.

რამდენიმე ეტაპიანმა მონაცემთა დამუშავებამ, განსაკუთრებით „stop_words“ პარამეტრის ოპტიმიზაციამ, „max_df“ და „n_components“ (თემის მოდელირებაში თემების წინასწარ განსაზღვრული რიცხვი) გაუმჯობესებასთან ერთად, უამრავი საათისა და ასეულობით გეგაბაიტი მეხსიერების დანახარჯის შემდეგ, სასურველ შედეგი გამოიღო. საბოლოოდ, თემის მოდელირების გაშვებულმა ალგორითმებმა დამაკმაყოფილებელი შედეგი დადეს. მიღებული კლასტერული სეგრეგაციის გრაფიკიც (იხ. ნახაზი 1) და სხვადასხვა თემებში ყველაზე ხშირად გამოყენებული სიტყვების მნიშნელობებიც მეტად საინტერესო გახდა.



ნახ.1. კლასტერული სეგრეგაციის გრაფიკი სწორი მონაცემებით

ნახაზი 1-ზე ხედავთ კლასტერულ სეგრეგაციას, რომელიც ორივე მეთოდისთვის საკმაოდ კარგად გამოიყურება. განსაკუთრებით, აღსანიშნავია NMF (მარჯვნივ) მეთოდის შედეგად მიღებული სეგრეგაცია. შემდეგ ნახაზზე წარმოდგენილია 30 გამორჩეული სიტყვა განაწილების მოდელიდან (ნახაზი 2).



ნახ.2. LDA მეთოდით მიღებული 30 ყველაზე მნიშვნელოვანი სიტყვა

როგორც ნახაზი 2-ზე ჩანს, გამორჩეულ სიტყვათა სიაში მოხვდნენ ისეთი სიტყვები, როგორებიცაა „კვანტური“, „სპინი“, „მასა“, „გალაქტიკა“, „ვარსკვლავი“, „სისტემები“, „ენერგია“ და სხვა. მიღებული შედეგიდან ჩანს, რომ ამ სიტყვათა უმეტესობა გამოსადეგია იმისთვის, რომ სავარაუდო სამეცნიერო მიმართულებას წარმოადგენდეს.

იმის შესაბამისად, რომ მონაცემები დაყოფილია ორ ნაწილად (ნაშრომის რეზიუმე და სრული ტექსტი) და გამოყენებულია დამუშავების ორი განსხვავებული მეთოდი, ალგორითმების გაშვების შემდეგ, მიღებული იქნა ოთხი მოდელი. ეს მოდელებია: LDA - რეზიუმეებისთვის, NMF - რეზიუმეებისთვის, LDA - სრული ტექსტისთვის და NMF - სრული ტექსტისთვის. ახლა, კი საჭიროა გამოვლენილ იქნას ორი საუკეთესო მოდელი რეზიუმესა და სრული ტექსტისათვის. ამისთვის, მოდელიდან ამოღებული იქნა ოთხივე მოდელში მიღებული, 20-20 თემის, 50-50 ყველაზე მნიშვნელოვანი

სიტყვა (ჯამში 80 თემა და 4000 სიტყვა). ამოღებული მონაცემები, თითოეული მეთოდისათვის სათითაოდ იქნა განხილული და შედეგად მიღებული იქნა ინფორმაცია კონკრეტული კლასტერის სასარგებლოდ გამოყენების მხრივ, პროცენტულად რომელ მეთოდს აქვს უკეთესი შედეგი. შედეგები გადმოცემულია 1-ელ ცხრილში.

თემის მოდელირების შედეგები, პოტენციური გამოყენების მხრივ ცხრ.2

Abstract NMF	Abstract LDA
≈0,9	≈0.95
Full text NMF	Full text LDA
≈60	≈67

როგორც ცხრილზე ჩანს, LDA მეთოდის გამოყენებამ ორივე, რეზიუმეებისა და სრული ტექსტების გამოყენების დროს, მცირედით აჯობა NMF მეთოდს. მიღებული თემების 50-50-ი ყველაზე რელევანტური სიტყვების გაანალიზების დროს აღმოჩნდა, რომ LDA მეთოდის მიერ კლასტერიზებული სიტყვები, NMF-თან შედარებით მცირედით უფრო გამოსადეგია და მეტად მიუთითებს რომელიმე კატეგორიასა თუ კვლევის მეთოდზე.

შემდეგი მანქანური სწავლების მეთოდი, რომელიც სადოქტორო კვლევის პროცესში იქნა გამოყენებული, ტექსტის კლასიფიკაციაა. აღნიშნული მეთოდის გამოყენებით მიღებულ იქნა სადოქტორო ნაშრომისთვის, ძალიან მნიშნელოვანი, კიდევ ერთ ინოვაციური სამეცნიერო სიახლე. კლასიფიკაციის მეთოდის გამოყენებისთვის არსებული სამეცნიერო ნაშრომთა მონაცემთა ბაზის მხოლოდ ერთი ნაწილი რეზიუმეები იქნა გამოყენებული. მიზეზი კი მისი პრაქტიკული გამოყენება გახლდათ.

ისევე როგორც თემის მოდელირების შემთხვევაში, კლასიფიკაციის ალგორითმების გაშვებამდე, საჭიროა არსებულ მონაცემთა წინასწარი დამუშავება, თუმცა მისგან განსხვავებით, ტექსტის კლასიფიკაცია

კონტროლირებადი მანქანური სწავლების მეთოდია. შესაბამისად, მას ორი შემავალი პარამეტრი გააჩნია. ეს პარამეტრებია თავად შესასწავლი ტექსტი და მისი კატეგორია. სადოქტორო კვლევის შემთხვევაში, დასახული მიზნებიდან გამომდინარე გამოყენებული იქნა არა კვლევის პროცესში გენერირებული, არამედ მონაცემთა მწარმოებლის მიერ მინიჭებული კატეგორიები. მრავალი კატეგორიის მქონე სტატიების შემთხვევაში არჩეულ იქნა პირველი კატეგორია. ამ მიზეზით არსებული ხარვეზები, გადაჭრილ იქნა მონაცემთა სიმრავლის მეშვეობით. საბოლოო შედეგი საკმაოდ დამაკმაყოფილებელი გამოდგა. შესწავლილი კატეგორიების ჯამური რაოდენობა ჯამში 178 გახლდათ. მას შემდეგ რაც მონაცემთა ბაზამ პირველად ელემენტარული დამუშავება გაიარა ის გასუფთავდა ზედმეტი მონაცემებისგან და მოხდა მისი მაქსიმალური შემცირება. ამის შემდეგ მოხდა მისი მომზადება კლასიფიკაციის ალგორითმების გასაშვებად. ამისთვის, პირველ ეტაპზე მონაცემები ორ ნაწილად იქნა დაყოფილი (იხ. ნახაზი 3).

```
[2]
from sklearn.model_selection import train_test_split

X = df['abstract']
y = df['category']

X_train, X_test, y_train, y_test=train_test_split(X, y,test_size=0.20,random_state=42)
```
Python

ნახ. 3. კლასიფიკაციის შესასწავლ მონაცემთა ორ ნაწილად დაყოფა

როგორც ნახაზი 3-ზე ხედავთ, წინასწარ მომზადებულია, მანქანური სწავლისათვის საჭირო ორი, რეზიუმეებისა და მათი კატეგორიების x და y მონაცემთა სიმრავლეები. ტექსტის კლასიფიკაციის მოდელების გასაშვებად მონაცემები ორ ნაწილად იქნა დაყოფილი. პირველი ნაწილი მოდელის შესასწავლად იქნა გამოყენებული, მეორე კი მისი ტესტირებისთვის. „sklearn” ბიბლიოთეკის model_selection მოდულიდან იმპორტირებული train_test_split

მეთოდის გამოყენებით, კი ამ მეთოდებისგან მიღებულია ოთხი X_train, X_test, y_train, y_test სიმრავლე. X_train და y_train სიმრავლეები, შესაბამისად, მონაცემთა შესასწავლად საჭირო რეზიუმეებსა და კატეგორიებს მოიცავს, ხოლო X_test და y_test მათი ტესტირებისთვის საჭირო რეზიუმეებსა და კატეგორიებს. როგორც ნახაზი 3-ზე ხედავთ, train_test_split მეთოდს, კიდევ ორი შემავალი პარამეტრი გააჩნია. test_size შედეგების დასატესტი სრულ მონაცემთა ნაწილის პროცენტს გულისხმობს. როგორც ხედავთ ამ პარამეტრის მნიშვნელობა 0.20-ია რაც იმას ნიშნავს, რომ კვლევის შემთხვევაში 1948654 სტატიის რეზიუმეს დამუშავებისას, შესასწავლი მონაცემების რაოდენობა 1558923-ია, ხოლო სატესტის მონაცემებისა 389731. კიდევ ერთ პარამეტრი, რომელსაც მეთოდის გამოყენებისას ხედავთ random_state-ია. აღნიშნული პარამეტრი გამოყენებული იქნა იმისთვის, რომ პარამეტრების ოპტიმიზაციის პროცესში, რამდენიმე განსხვავებული მანქანური სწავლების მეთოდების გაშვების დროს სასწავლო და სატესტი მონაცემების გადანაწილება განსხვავებულად არ მომხდარიყო, რასაც შემდეგ, რეზულტატის შედარებების დროს შეცდომაში შევეყვანა შეეძლო გამოეწვია. ისევე როგორც თემის მოდელირებისას, კლასიფიკაციის შემთხვევაშიც, დოკუმენტთა ვექტორიზაცია მონაცემთა წინასწარი დამუშავების ერთ-ერთი მნიშვნელოვანი კომპონენტია. ზემოთ აღნიშნული თითოეული მეთოდისთვის ვექტორიზაციის სხვადასხვა მეთოდი იქნა გამოყენებული. საბოლოოდ, მოდელების მიღების შემდეგ კი „predict“ მეთოდის გამოყენებით, მოდელში წინასწარ გადადებული სატესტო მონაცემების გაშვება მოხდა. მიღებული ინფორმაცია კი გამოყენებული იქნა იმისთვის, რომ მოგვეხდინა მოდელის სიზუსტის მაჩვენებელის განსაზღვრა, რაც მის საბოლოო ხარისხზე მიუთითებს. საბოლოოდ მოხდა ორივე მეთოდის გამოყენების შემდეგ მიღებული სიზუსტის მაჩვენებელი ქულების ერთმანეთთან შედარება და მიღებული იქნა საბოლოო მოდელი.

იმ მიზეზით, რომ TF-IDF ვექტორიზერი, ავტომატურად ვექტორიზაციის

დროს სიტყვას მნიშნელობას 0-დან 1-მდე ანიჭებს, ამ სიტყვის ყველა დოკუმენტში გამოყენებადობის გათვალისწინებით, მეთოდების გაშვება მონაცემთა წინასწარი დამუშავების სამი სხვადასხვა გზის მიხედვით მოხდა. ეს გზები იყო: 1 - მონაცემთა ვექტორიზაცია TF-IDF მეთოდის გამოყენებით დამატებითი პარამეტრების გარეშე. 2 - მონაცემთა ვექტორიზაცია TF-IDF მეთოდის გამოყენებით დამატებითი პარამეტრების გამოყენებით - max_df=0.5, min_df=2. და 3 - მონაცემთა ვექტორიზაცია CountVectorizer მეთოდის გამოყენებით დამატებითი პარამეტრებით - max_df=0.5, min_df=2. პირველი მეთოდი, რომელიც სტატიების რეზიუმეთა კლასიფიკაციისათვის გაეშვა, მიამიტი ბეიესის ალგორითმია, ხოლო მეორე მეთოდი ხაზოვანი მხარდაჭერის ვექტორული კლასიფიკატორი SVC-ია. საბოლოო შედეგები ასე გამოიყურება.

ტექსტის კლასიფიკაციის მეთოდების შედარება სხვადასხვა ვექტორიზაციის გამოყენებით **ცხრ.2**

	მიამიტი ბეიესი	SVC
CountVectorizer პარამეტრებით	0.54	0.51
TF-IDF	0.36	0.593
TF-IDF პარამეტრებით	0.41	0.591

როგორც ცხრილში ხედავთ მიამიტი ბეიესის ალგორითმი უკეთესს შედეგს აჩვენებს CountVectorizer-ის გამოყენების დროს, მაგრამ არადამაკმაყოფილებელია TF-IDF-ის შემთხვევაში. ხოლო SVC ორივე TF-IDF მეთოდის გამოყენების დროს დამაკმაყოფილებელ შედეგს გვაძლევს. შედარების შემდეგ გამოვლინდა საბოლოო გამარჯვებული (მწვანე ფერში). SVC მეთოდითა და პარამეტრების გარეშე გაშვებული TF-IDF ვექტორიზერით მიღებულ იქნა ინოვაციური მოდელი, რომელიც სამეცნიერო ნაშრომთა 200-მდე კატეგორიაში, 60%-იანი სიზუსტის მაჩვენებელს აფიქსირებს.

5. ტექსტური ინფორმაციის შექმნისათვის საჭირო ფუნქციონალისა და ინტერფეისი კვლევა და მისი შედეგები

მოცემულ თავში საუბარია კვლევის შემდეგ ორ ეტაპზე. ესენია: კონტექსტუალური გამოკითხვები და სხვადასხვა ტიპის სამეცნიერო ნაშრომებში არსებული ფორმების კვლევა და შედეგები. მეორე მნიშვნელოვანი გამოწვევა, რომლის გაუმჯობესების მიზნად დასახვის გამოც მოხდა აღნიშნული გამოკითხვის ჩატარება, ტექსტურ ინფორმაციაზე მომუშავე რედაქტორის ფუნქციონალის ნაკლოვანებებისა და მისი არასრულყოფილების გამოვლენაა. გამოკითხვისას დასმული შეკთხვები, ეხებოდა რესპოდენტთა მიერ გამოყენებულ ტექსტურ რედაქტორებს: დაისვა შეკითხვები იმასთან დაკავშირებთ თუ რა მოსწონდათ და არ მოსწონდათ მათ ამ რედაქტორებში, რას სრულყოფდნენ და რის დანაკლისს განიცდიდნენ, რა დამატებით ინსტრუმენ-ტებს იყენებდნენ წერის პროცესში ტექსტური რედაქტორის გარდა და რაში.

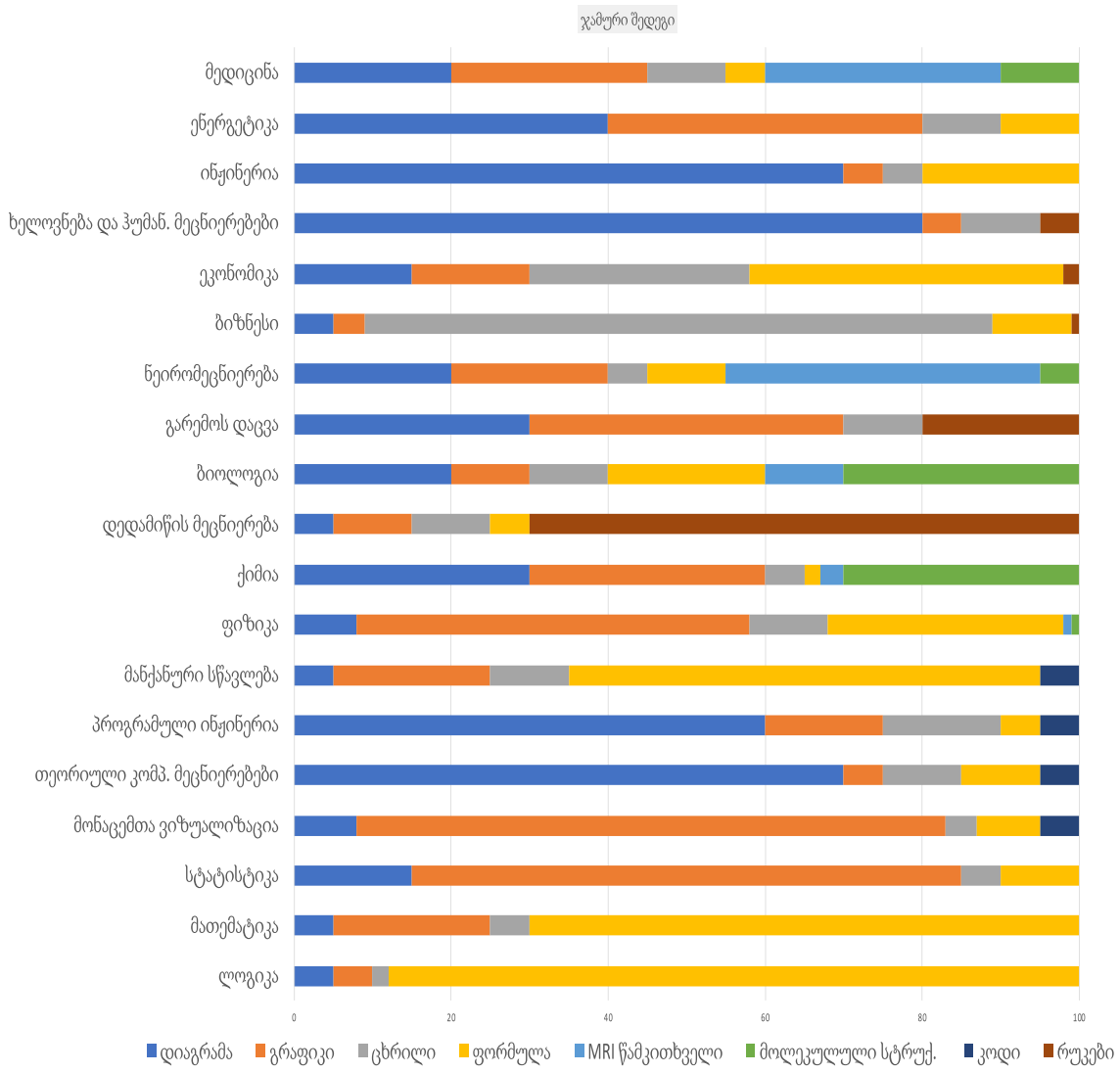
ამასთან, ზოგიერთი შეკითხვა, დოკუმენტებთან მუშაობის ბიზნეს პროცესში არსებული ხარვეზების აღმოსაჩენად იყო დასმული. გამოკითხვაში მონაწილე პირებმა პასუხი გასცეს შეკითხვებს თუ როგორ ახდენდნენ ისინი ნაშრომთა შენახვასა და გამოქვეყნებას, როგორ ახდენდნენ თანაავტორებთან ურთიერთობასა და მათ მიერ ნაშრომის განხილვის პროცესის მენეჯმენტს. გამოვლენილი იქნა, რომ არსებობს თანაავტორებს შორის ნაშრომის განხილვის სამი ძირითადი გზა. ეს მეთოდები სტანდარტულია, თუმცა, არ არსებობს ამ გზების ავტომატიზების არსებული გადაწყვეტები. როგორც აღმოჩნდა თანაავტორებთან ურთიერთობის სამი ძირითადი მიდგომა არსებობს. აღვწეროთ თითოეული მათგანი: წარმოვიდგინოთ, რომ სამეცნიერო სტატიას ორი თანაავტორი ყავს. როდესაც გამზადდება ნაშრომის პირველი ვერსია და უკვე დროა, რომ ის თანაავტორებმა განიხილონ, პირველ რიგში, სისტემაში უნდა იქნას მათი ელ-ფოსტის მისამართების შეყვანა. ამის შემდეგ უნდა მოხდეს ნაშრომის განხილვის სამი შესაძლო პროცესიდან ერთ-ერთის არჩევა და პროცესის დაწყება. პირველი მეთოდის არჩევის შემთხვევაში, ნაშრომი ჯერ გადაეგზავნება რიგით პირველ თანაავტორს, მისი განხილვის, კომენტარებისა

და შენიშვნების ჩამატების შემდეგ კი გადაეგზავნება რიგით მეორე თანაავტორს და მხოლოდ მას შემდეგ რაც მეორე ავტორიც დაასრულებს განხილვას, დაბრუნდება პირველ ავტორთან. მეორე მეთოდის დროს ნაშრომი ჯერ რიგით პირველ თანაავტორს გადაეგზავნება, მისი განხილვის შემდეგ კი, განსხვავებით პირველი მეთოდისაგან, კვლავ პირველ ავტორს დაუბრუნდება. მას შემდეგ, რაც უკვე პირელი ავტორი შეიტანს სასურველ ცვლილებებს სისტემაში და მიუთთებს, რომ ნაშრომი მზადაა შემდეგი განხილვისათვის, ის მეორე თანაავტორს გადაეგზავნება. რაც შეეხება მესამე მეთოდს, ეს მეთოდი ძირითადად ისეთ შემთხვევებში გამოიყენება როდესაც ნაშრომის განხილვისას დროის სიმწირესთან გვაქვს საქმე. მოცემულ მეთოდში, როდესაც ნაშრომი განსახილველად იგზავნება, ის პარალელურად გაეგზავნება ორივე თანაავტორს.

შემდეგი კვლევა რომელზეც მეხუთე პარაგრაფშია საუბარი, სხვადასხვა ტიპის სამეცნიერო ნაშრომებში არსებული ფორმების კვლევა და მისი შედეგებია. განხილვისას, თითოეული ნაშრომისთვის ამოწერილი იქნა მასში გამოყენებული განსხვავებული ფორმების სია. ყურადღება მივანიჭეთ ისეთი ფორმების არსებობას, რომელთა გამოყენებისთვისაც რაიმე სპეციალური ფუნქციების არსებობაა საჭირო. ასეთ ფორმებად აღებულია შემდეგი: დიაგრამები, გრაფიკები, ცხრილები, ფორმულები, სამედიცინო სურათები, მოლეკულური სტრუქტურები, პროგრამული კოდები და რუკები. ამასთან ერთად, ყურადღება იქნა გამახვილებული კონკრეტულად რა ტიპის გრაფიკები და დიაგრამებია გამოყენებული ამ ნაშრომებში (მაგალითად კლასების დიაგრამა, BPMN დიაგრამა, სხვადასხვა ტიპის გრაფიკები და სხვა). აგრეთვე ყურადღება იქნა გამახვილებული ნაშრომებში გამოყენებულ სიმბოლოებზე, რადგან გამოკითხვის შედეგად, მიღებული კომენტარებიდან, არაერთი იყო იმასთან დაკავშირებით, რომ ნაშრომზე მუშაობისას, არსებულ ტექსტურ რედაქტორებში, საკმაოდ რთულია სასურველი სიმბოლოების მოძებნა. სწორედ

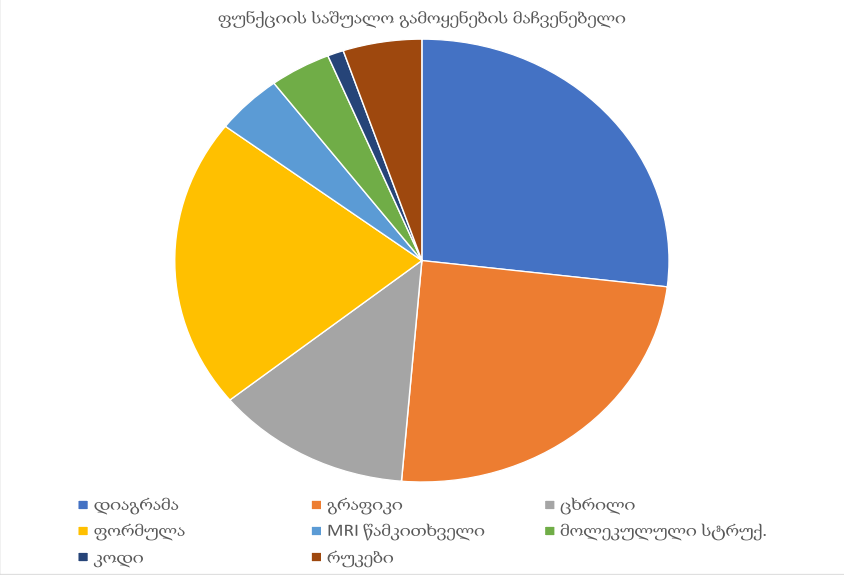
ამიტომ, ისეთ სამეცნიერო მიმართულებებში სადაც ხშირადაა გამოყენებული სიმბოლოები, ჩანიშნული იქნა ისინი. მაგალითად, მათემატიკურ სიმბოლოებში, ჭარბობს კალკულაციის, ტოლობის და სხვა მსგავსი სიმბოლოები, ხოლო ლოგიკაში ხშირია - თანაკვეთა, გაერთიანება და სხვა. დავუბრუნდეთ ფორმებს და წარმოვადგინოთ იმის პროცენტული მაჩვენებელი, თუ რამდენადაა თითოეული ფორმა გამოყენებული განსხვავებულ სამეცნიერო თემებში.

როგორც მე-4 ნახაზიდან ჩანს, სხვადასხვა კატეგორიის სამეცნიერო ნაშრომებში გამოყენებული ფუნქციონალი საკმაოდ მრავალფეროვანია. თუ განვიხილავთ თითოეული მიმართლების გამოყენებულ ფორმებს, აღმოვაჩენთ, რომ სხვადასხვა ტიპის ნაშრომში სხვადასხვა ფორმებია პრიორიტეტული. მაგალითად, ლოგიკასა და მათემატიკაში, ფორმების უმრავლესობა ფორმულებია, როდესაც სტატისტიკასა და მონაცემთა მეცნიერებაში გრაფიკები ჭარბობს.



ნახ.4. ერთიანი გამოყენებული ფორმების მაჩვენებელი თითოეული სამეცნიერო მიმართულებისთვის

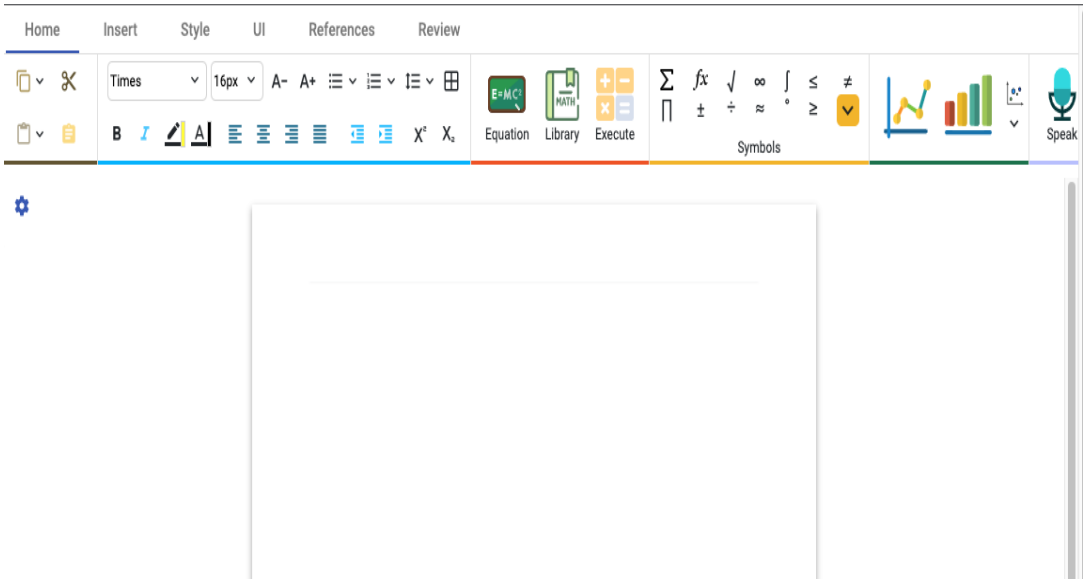
წარმოვადგინოთ გრაფიკი, სადაც გამოჩნდება ჯამურად რა რაოდენობის ფორმებია გამოყენებული ყველა ტიპის სამეცნიერო ნაშრომებში (ნახაზი 5).



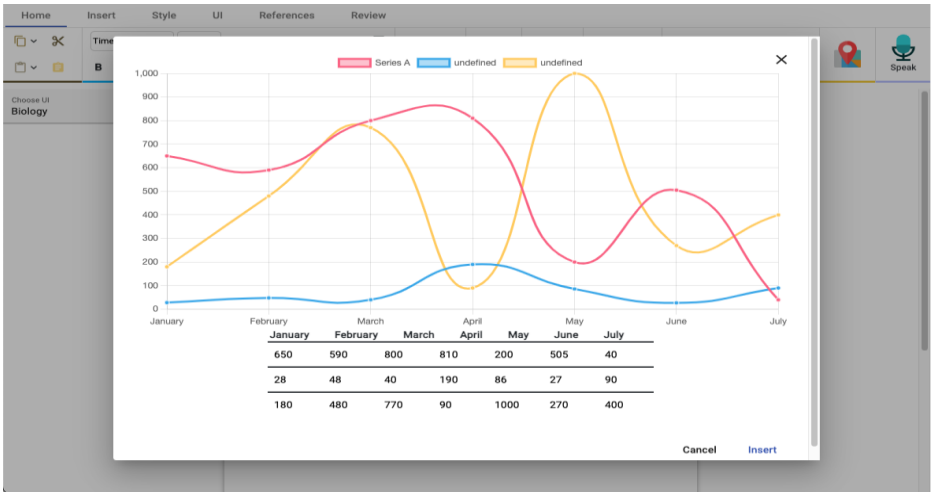
ნახ. 5. სამეცნიერო ნაშრომებში ჯამურად გამოყენებული ფორმები

6. პროტოტიპი

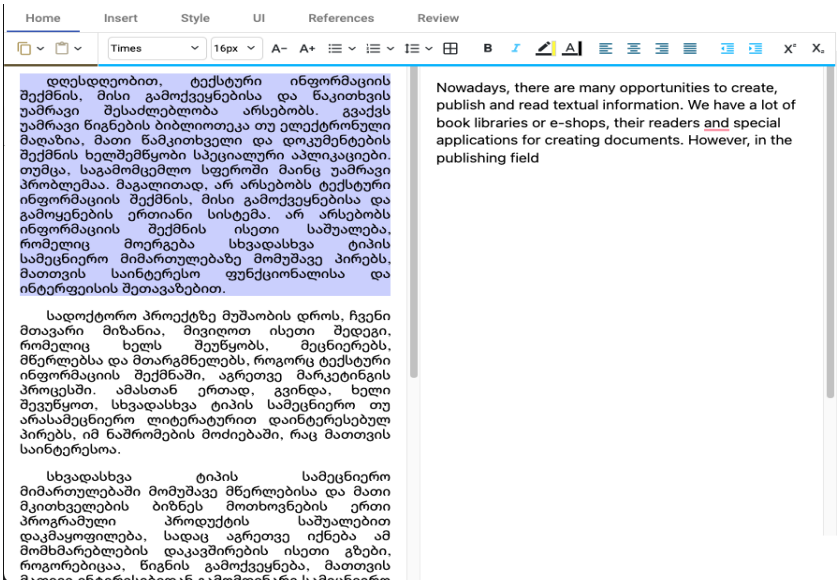
მეექვსე პარაგრაფში, ნაშრომში ჩატარებული კვლევებიდან გამომდინარე წარმოდგენილია, შექმნილი სისტემის პროტოტიპი, რომელიც სხვადასხვა სფეროში მომუშავე პირებს, სპეციალურად მათთვის მორგებულ მომხმარებლის ინტერფეისსა და ფუნქციებს შესთავაზებს (იხ. ნახაზი 6, 7 და 8).



ნახ. 6. მომხმარებლის ინტერფეისი მათემატიკური ნაშრომისთვის



ნახ. 7. Carts.js მუშაობის პროცესი დიალოგურ ფანჯარაში



ნახ. 8. სპეციალურად მთარგმენთათვის შექმნილი მომხმარებლის ინტერფეისი

პარაგრაფში ასევე ვისაუბრეთ სისტემის უსაფრთხოებაზე, ინდენთიფიკაციისა და ავტორიზაციის გამოყენებით.

ზოგადი დასკვნები

ჩვენთვის საინტერესო კატეგორიის, სამეცნიერო ნაშრომების პოვნა, რომელიც ჩვენთვის ასევე საინტერესო მეთოდებითაა შესრულებული, მარტივი პროცესი არაა. მანქანური სწავლების, თემის მოდელირების ალგორითმების

გამოყენებით მიღებული იქნა, ორი ინოვაციური მოდელი, რეზიუმესა და სრული ტექსტისთვის, რომელთა სინქრონულად გამოყენების შემთხვევაშიც, რომელიმე სამეცნიერო ნაშრომით დაინტერესებული ადამიანისთვის, როგორც თემატურად, ასავე მეთოდოლოგიურად საინტერესო ნაშრომების შეთავაზება ხდება შესაძლებელი.

მანქანური სწავლების, ტექსტის კლასიფიკაციის მეთოდით, მიღებული იქნა ინოვაციური მოდელი მაღალი სიზუსტის მაჩვენებლით, რისი გამოყენებითაც, ტექსტური ინფორმაციის შემქმნელ ადამიანს, მისი დაწერილი ტექსტიდან გამომდინარე, მისივე სამუშაო მიმართულებით ხშირად გამოყენებულ ფუნქციონალსა და მომხმარებლის ინტერფეისს შესთავაზებს.

დღესდღეობით არსებულ ყველაზე პოპულარულ ტექსტურ რედაქტორებს სამეცნერო ნაშრომის შექმნისათვის მნიშვნელოვანი ფუნქციონალი არ გააჩნია. გამოკითხვის შედეგად, გაანალიზებული იქნა ინფორმაცია ამ ფუნქციონალის საჭიროებებზე, შემდეგ კი გამოკვლეულ იქნა ამ ფუნქციონალის შესრულების მეთოდები.

საგამომცემლო სფეროში არსებული ბიზნეს პროცესები, მათი შესრულების გამარტივების მიზნით, ავტომატიზაციას საჭიროებს. გამოკითხვების შედეგად ჩამოყალიბებული იქნა ამ პროცესების ავტომატიზაციის მოდელები.

დღევანდელი ტექსტური რედაქტორები, არაა მორგებული კონკრეტულ სამეცნიერო სფეროზე მომუშავე პირების საჭიროებებზე. გამოკვლეული იქნა სხვადასხვა კატეგორიის მქონე სამეცნიერო ნაშრომებში გამოყენებული ფუნქციებისა და ფორმების რაოდენობრივი მაჩვენებლები. შედეგად კი, შექმნილი იქნა პროტოტიპი, რომელიც სხვადასხვა სამეცნიერო სფეროში მომუშავე ტექსტური ინფორმაციის შემქმნელ ადამიანებს, სპეციალურად მათზე მორგებულ სამუშაო გარემოს სთავაზობს.

ნაშრომის აპრობაცია

სადისერტაციო ნაშრომის ძირითადი შინაარსი მოხსენებულ იქნა ინფორმატიკისა და მართვის სისტემების ფაკულტეტის „მართვის ავტომატიზებული სისტემების (პროგრამული ინჟინერია)“ კოლეგიის სამეცნიერო სემინარებისა და კოლოქვიუმების სახით. კვლევები მოხსენიებულ იქნა კონფერენციაზე:

- საინფორმაციო საზოგადოება და განათლების ინტენსიფიკაციის ტექნოლოგიები (ISITE ’21). თემის სათაური: MODELING OF BUSINESS PROCESSES FOR THE COMBINED SYSTEM OF PUBLISHING MARKETING AND CREATION OF TEXTUAL INFORMATION. მოხსენების თარიღია: 21.05.2021.

კვლევებთან დაკავშირებული შედეგები გამოქვეყნებულია სტატიებში:

1. გოგშელიძე დ., სურგულაძე გ., თურქია ე. ტექსტური ინფორმაციის შექმნისა და საგამომცემლო მარკეტინგის ერთიანი სისტემის ბიზნესპროცესების მოდელირება. სტუ-ს შრ.კრ., „მართვის ავტომატიზებული სისტემები“, No1(25), **2018**, 41-48. EISSN 1512-2174,

2. გოგშელიძე დ., RESEARCH OBJECTIVES AND METHODOLOGIES, WHILE PROCCESSING OF “UNITED SYSTEMS OF CREATING TEXTUAL INFORMATION AND PUBLISHING MARKETING”. სტუ-ს შრ.კრ., „მართვის ავტომატიზებული სისტემები“, No2(26), **2016**, 261-264. EISSN 1512-2174,

3. გოგშელიძე დ., გოგიშვილი ა. სხვადასხვა კომპონენტების განსაზღვრა ტექსტური ინფორმაციის შექმნისა და საგამომცემლო მარკეტინგის ბიზნეს პროცესების მოდელირებისას. სტუ-ს შრ.კრ., „მართვის ავტომატიზებული სისტემები“, **2019**, No 2(29), 178-182. EISSN 1512-2174,

4. გოგშელიძე დ. ტექსტური ინფორმაციის შექმნის ბიზნესპროცესების მოდელირების მხარდამჭერი პლატფორმა სერვის ორიენტირებული მიდგომით. სტუ_ს შრ.კრ., „მართვის ავტომატიზებული სისტემები“, **2022**, No 1(33), vol.2. 41-44. EISSN 1512-2174, DOI.org/10.36073/1512-3979

ABSTRACT

Nowadays, there are many opportunities to create, publish and read textual information. We have a lot of electronic libraries, e-book shops, book readers and applications for creating documents. However, in the publishing field, there is still a need for development. For example, there is no single system for creating, publishing, and using textual information. There is no way to create information that suits people working in different types of science by offering them functionality and interface designed especially for their working style and requirements.

While working on a doctoral project, our main goal is to achieve a result that will help scholars, writers and translators both in the creation of textual information and in the process of publishing it. In addition, we want to help people interested in different types of scientific or non-scientific literature to find papers and works that fit their field of interest.

Meeting the business needs of writers and their readers through a single software product, which will also include ways to connect them, publish a book, offer them scientific papers or other types of textual information based on their interests, specially customized interface for writers working on different types of scholarly work, etc, allows us to create innovative approaches, on our way to model a business process in our project. Among other things, the most important innovative aspect of the project is, around which we have both scientific news and very important practical values - The mentioned system of facilitating the creation of scientific papers.

Various types of research have been conducted in the paper, we have studied millions of scientific papers in order to create a machine learning model through which the system, at the beginning of the paperwork, automatically assigns it to the appropriate category. After that, we studied the methodologies of writers' work and their approaches to writing. we studied the type of functionality required when working on each category of scientific papers, and we created a different user interface for each different case.

Regarding the practical application of the system designed during the doctoral dissertation, we can say that it meets the requirements of life, such as:

- It will be possible to write a scientific paper with a user interface and functionality created specifically for the scientific field of the writer, which will make the writing process simpler and more multifunctional.
- The system brings together textual information for both the workers (its creators and translators) as well as its users, which means a direct

connection between the writer and the reader. This underscores the innovative side of the project.

- Thanks to improved processes, scientists working with textual information, due to their simplicity and multifunctional capabilities, will be able to save time, which will have a positive impact on the scientific quality of the paper.
- Translators, through the project, will be able to translate more easily through the interface created specifically for them, and at the same time, they will be able to directly contact the authors of the translated text.

In the paper, we have conducted research in various directions. Among the results are models obtained by machine learning's Topic Modeling algorithms, Text Classification Models, business processes described as a result of customer surveys, information obtained from surveys about the existing problems while working on the textual information, information obtained from the review of different categories of scientific papers about the different functions and forms used in each of them, and more.

- Therefore, the research process is divided into four main parts. these are:
- Modeling the topics according to scientific papers.
- Classification of scientific papers by studying their categories.
- Survey of scientists. Research of business processes of the required functionality and workflow.
- Review different categories of scientific papers and describe the functionality used in the fields.

After that, we created a prototype with examples of its creation and some details. As already mentioned, the aim of doctoral research is to simplify the process of working with textual information and with this, improve the work experience of its creator. Also, through the machine learning's Topic Modeling models, we can say that it will help scientific workers to popularize their works.