

ExploreGeorgia

მონაცემთა ანალიზი და მანქანური სწავლება

დემეტრე ბეროშვილი

Email: demetre.beroshvili512@ens.tsu.ge demetre.beroshvili@gmail.com

რეზიუმე – ExploreGeorgia წარმოადგენს საქართველოში მოგზაურობის ვებ-საიტს, რომელიც მომხმარებლებს საშუალებას მისცემს იოლად დაგეგმონ თავიანთი მოგზაურობა მათთვის მიწოდებული რეკომენდაციებითა და რჩევებით და შემდეგ დაჯავშნონ მოგზაურობის ბილეთები და სასტუმროების ნომრები. ჩვენი პროექტი მოიცავს მოსახერხებელი ინტერფეისს, რომელიც საშუალებას აძლევს მომხმარებლებს ეფექტურად დაგეგმონ მოგზაურობები. შესაძლებლობა მიცემთ გადაწყვიტონ სად რა აქტივობა იქნება უმჯობესი მათთვის, იქნება ეს ლაშქრობა, ღირშესანიშნაობების დათვალიერება, მუზეუმების ნახვა თუ სხვა. საქართველოს მრავალფეროვანი ბუნება, მოგზაურობის მსურველების დიდ არჩევანს აძლევს მოხვდნენ ნებისმიერი არსებული ჰავისა და კლიმატის კუთხეებში. ხოლო, საქართველოს მდიდარი კულტურაც ჰმატებს ყოველ კუთხეს თავის ხილბს.

ამდენი მრავალფეროვანი და კომბინაციური არჩევანის გასაკეთებლად საჭიროა პროგრამული მხარდაჭერა, რომელიც მოიცავს რეკომენდაციის სისტემას. შესაბამისი სერვისების შესაქმნელად საჭიროა მომხმარებლების მიერ შექმნილი მონაცემების სტატისტიკური შესწავლა. ამ დარგისთვის მონაცემების დამუშავება და შესწავლა კარგად შესწავლილია და დამუშავებული ინფორმაციის თეორიისა და მანქანური სწავლების დარგებში. მაღალშედეგიანი შეთავაზებების გამოთვლის ალგორითმები და სისტემები უკვე კარგად არის გამოკვლეული მრავალი ინდუსტრიის მიერ, მაგალითად, კომპანები, როგორებიცაა Netflix და Spotify იყენებენ შესაფერის სისტემებს, რომლებიც აკვირდება მომხმარებლების გემოვნებას და შემდეგ იძლევა რეკომენდაციებს. მსგავსი სისტემებია დანერგილი ასევე ონლაინ-მაღაზიებში, რათა მომხმარებლებს უფრო ზუსტი და შესაფერისი პროდუქტები წარუდგინონ. ადგილის არჩევა და ცუდი არჩევანის გაკეთების შემთხვევაში მთელი მოგზაურობა ეშლებათ მოხალისეებს. ჩვენი ვებ-გვერდის მიზანია ამ პრობლემის აღმოფხვრა და გაუთვალისწინებელი პრობლემების მინიმუმამდე დაყვანა.

ExploreGeorgia არის დინამიური სისტემა, რომელიც მუდმივად ჩაბმულია და მორგებულია კემპინგის მოყვარულთა მრავალფეროვან საჭიროებებს. ჩვენი კოლექტიური ძალისხმევით ვცდილობთ შევქმნათ მყუდრო, სანდო გარემო, რომელიც ითვალისწინებს ინდივიდუალურ საჭიროებებს და ხელს უწყობს ძიების და თავგადასავლების გრძნობას.

გასაღები სიტყვები - დაგეგმვა, მონაცემთა დამუშავება, მომხმარებლების მეგობრულობა.



1 შესავალი

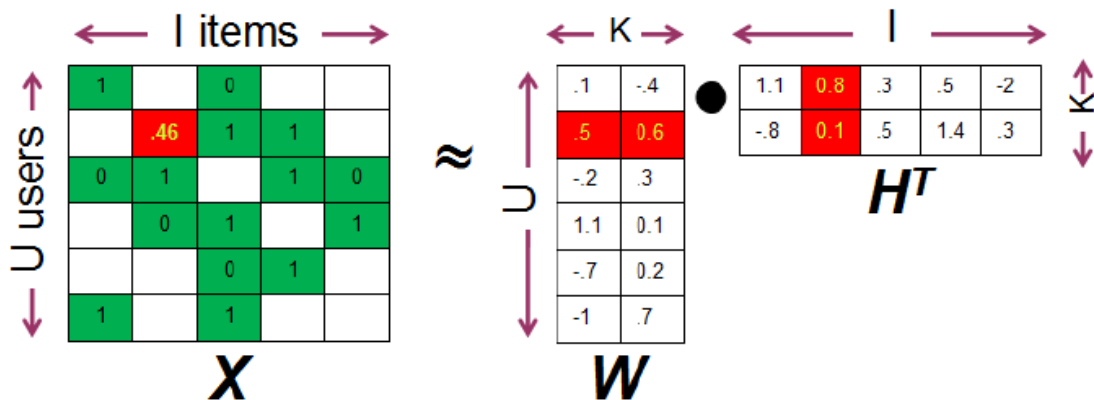
საქართველოს განვითარებული ტურიზმის ინდუსტრია და მრავალფეროვანი ბუნება და ღირშესანიშნაობები ქმნის საუკეთესო საშუალებას ასეთი პროექტის განსავითარებლად. ExploreGeorgia-ის განვითარების მრავალმხრივი ლანდშაფტი მოიცავს მონაცემთა ანალიტიკასა და მომხმარებლებზე ორიენტირებული რეკომენდაციების წარდგენას. ჩვენი ტექნიკური სფერო მოიცავს მძლავრ მონაცემთა ანალიზს, მანქანური სწავლებასა და შესაბამისი ალგორითმების გამოყენებას.

ჩვენი მოპოვებული მონაცემები დამუშავდება Python პროგრამირების ენაზე,

მონაცემების დამუშავებისთვის და ანალიზისთვის განკუთვნილი ბიბლიოთეკების გამოყენებით, როგორცაა Pandas, NumPy და ინფორმაციის ვიზუალიზაციისთვის Matplotlib-ს. ასევე, გამოყენებული იქნება მანქანური სწავლების ბიბლიოთეკები, როგორებიცაა Scikit-learn ან/და TensorFlow, რათა მოხდეს მომხმარებლების სრულყოფილი კლასტერიზაცია და ადაპტიური სწავლა. ზემოხსენებული მანქანური სწავლების ბიბლიოთეკები ასევე გამოყენებული იქნება მატრიცული ფაქტორიზაციისთვის. გარდა ამისა, ამ მოდელების შესაქმნელად, დასახვეწად და გუნდური თანამშრომლობისთვის ვიყენებთ GitHub პლატფორმას. ჩვენ გადავხედავთ მონაცემთა ვრცელ ნაკრებებს, რათა გამოვიტანოთ ყველა შესაძლო ნიუანსური დასკვნა მოგზაურებისა და მათი რჩეული ადგილების შესახებ. Python-ისთვის გამოყენებული იქნება Jupyter Notebook მონაცემების დამუშავებისა და სტატისტიკური მოდელირებისთვის, რადგან იგი იძლევა კოდის მონაკვეთების ცალ-ცალკე გამოძახების საშუალებას და ნელი, მძლავრი ოპერაციების ჩატარება მხოლოდ ერთხელ იქნება საჭირო.

2 მონაცემების შეგროვება და მათი დამუშავება,

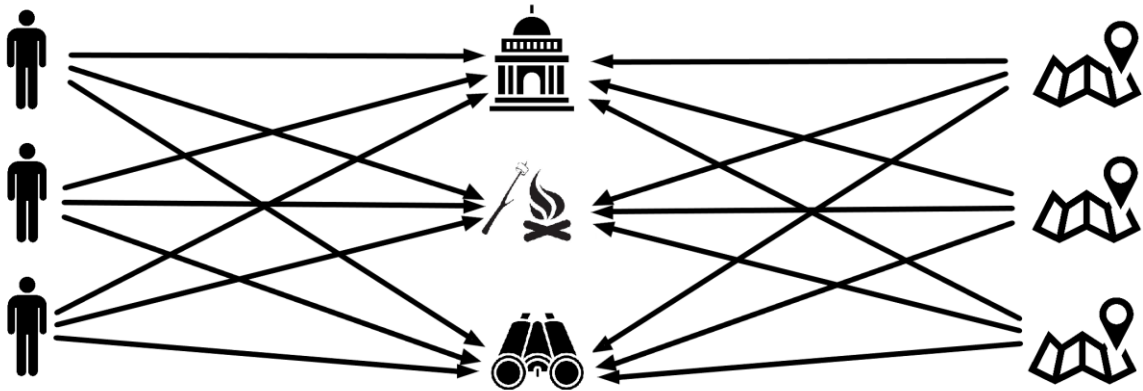
ინტერნეტის წარმოშობიდანვე მალე აქტუალური გახდა მონაცემების შეგროვება და დამუშავება. მონაცემების შეგროვების დროს გასათვალისწინებელია, რომ არ ირღვევა მომხმარებლების უფლებები. ExploreGeorgia-ს მიერ შეგროვებული მონაცემები არ არის მიზნული კონკრეტულ ინდივიდებზე, მხოლოდ დროებით, უსახელოდ ინახება მათი შეფასებები და რეაქციები, შემდეგ ეს შეფასებები მატრიცული ფაქტორიზაციით ერთმანეთს ერწყმის და რჩება მხოლოდ მისი ერთგვარი „ანასახი“, რომლის უკუქცევაც შეუძლებელია.



ამ მაგალითში 5x6 მატრიცა იყოფა ორ 2x6, და 6x2 მატრიცებად.

მომხმარებლებისთვის რეკომენდაციების გაწევა ერთ-ერთი ურთულესი პრობლემაა, რომლის გადაჭრასაც უკვე რამდენიმე დეკადაა კომპანიები ცდილობენ. შესაბამისად შექმნილ დარგს ეწოდება ინფორმაციული თეორია, ძირითადად გამოიყენება კოლაბორაციული ფილტრაცია (მომხმარებლების უპირატესობების შედარება) და კონტენტის ფილტრაცია (მომხმარებლების მონაცემებით განსაზღვრული). ორივე მეთოდს გააჩნია მრავალი ნაკლი და შესაბამისად, უკეთესი შედეგების მისაღებად, ჩვენი რეკომენდაციის სისტემა მოიცავს ორივე მეთოდის გაერთიანებულ ვერსიას. ასევე გამოყენებული იქნება მატრიცული ფაქტორიზაცია, რათა მონაცემები უკეთესად დამუშავდეს და მეხსიერების განთავისუფლება შევძლოთ. თუმცა, ამ მეთოდის გამოყენება საჭიროებს დიდი რაოდენობით მონაცემებს და მძლავრ მანქანურ სწავლებას. ასევე გასათვალისწინებელია, რომ ადამიანების გემოვნება და უპირატესობები იცვლება დროთა

განმავლობაში, ამიტომ შემთხვევითობის პირდაპირი გამოყენებით შესაძლებელია მომხმარებლები ახალ მოგზაურობის სტილს გავაცნოთ.



კოლაბორაციული ფილტრაციის მარტივი დიაგრამა, ყველა კავშირს გააჩნია განსაზღვრული წონა და შემდეგ ითვლება მომხმარებლებს, ინტერესებსა და ადგილმდებარეობებს შორის მსგავსების კოეფიციენტი.

ასევე ყურადსაღებია, რომ დასაწერია ამ ინფორმაციის მოძიების სისტემა, განსაკუთრებით კონტენტის ფილტრაციისთვის. ასეთი სარეკომენდაციო სისტემისთვის საჭიროა მომხმარებლების ერთგვარი „გამოკითხვა“, ეს შეიძლება იყოს შემოსახაზი კითხვების ნაკრები. თუმცა, ამ მეთოდს ძალიან ხშირად აკრიტიკებენ, რადგან მომხმარებლები კითხვების უმრავლესობას დაჩქარებით პასუხობენ, კითხვარის ბოლოს უფრო მობეზრებულები არიან და შეიძლება აღარც შეავსონ. ამიტომ უმჯობესია ინფორმაციის დაგროვების უფრო მარტივი, შეუმჩნეველი მეთოდები. მაგალითად, შესაძლებელია წარვადგინოთ რამდენიმე მარშრუტი ერთად და დავაკვირდეთ მომხმარებელი პირველად რომელს დააკვირდება, რომელს დააკლიკებს. ეს უფრო კარგად წარმოდგენას შეგვიქმნის მომხმარებლის ქვეცნობიერ სურვილებზე და გემოვნებაზე, რაც ხშირად უფრო სანდო და ხელსაყრელია.

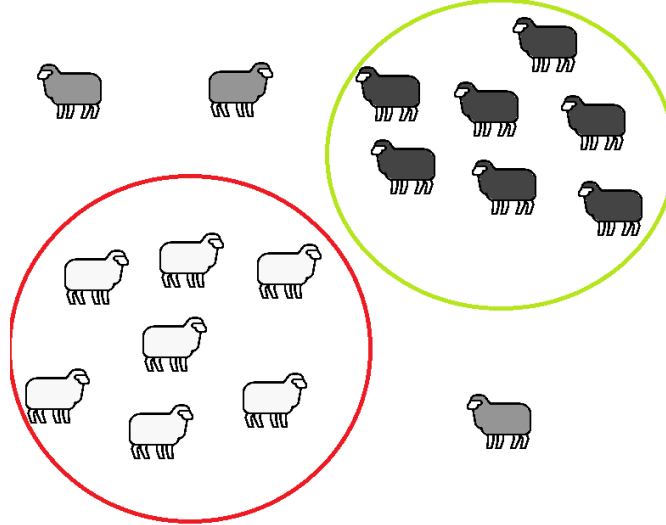
ამ ეტაპზე შექმნილია ვებ-გვერდის პირველადი „ალფა“ ვერსია არსებობს. თუმცა მას არ გააჩნია არანაირი ინფორმაციის შეგროვებისა და დამუშავების მექანიზმი. ეს სისტემა სრულიად ნულიდანაა შესაქმნელი. ზედაპირულად შესწავლილი არის ინფორმაციის თეორია, მანქანური სწავლების საკითხები. ასევე შერჩეულია ბიბლიოთეკები, მეთოდები და პროგრამული მხარდაჭერი სერვისები.

3 მოსალოდნელი პრობლემები

ერთ-ერთი მთავარი პრობლემა არის ტექნიკური რესურსები. მომხმარებლებისთვის ხელსაყრელი რჩევების მიცემას სჭირდება უზვი მონაცემები, მოგზაურობის სერვისისთვის კი ეს რთული მოსაპოვებელია, მომხმარებლები ხშირ შემთხვევაში წელიწადში მხოლოდ ერთხელ ან ორჯერ მოგზაურობენ. ასევე დგას მომხმარებლების მოგიზდვის რთული საკითხი. ამიტომ, პროგრამული უზრუნველყოფის გასატესტად გამოყენებული იქნება ხელით შედგენილი ყალბი, მაგრამ რეალურთან ახლოს მყოფი მონაცემები.

კოლაბორაციული და კონტენტური ფილტრაციისა და ასევე კლასტერიზაციის კიდევ ერთი პრობლემა არის, რომ ისინი ვერ ითვალისწინებენ ეგრედწოდებულ „ნაცრისფერ ცხვრებს“ (Grey sheep), ანუ განსაკუთრებულად უნიკალური გემოვნების მომხმარებლებს. მათი მონაცემებთან რაიმე საერთოს პოვნის საშუალება შეიძლება არ არსებობდეს და

შეუძლებელი ხდება მათთვის მორგებული რეკომენდაციების გაწევა. თუმცა, ჩვენდა საბედნიეროდ ეს პრობლემა მხოლოდ მომხმარებლების ძალიან მცირე ნაწილს ეხება და არ არის მეთოდის მთლიანად შეცვლისთვის საკმარისი საფუძველი.



4 ფორმულები

კოლაბორაციული ფილტრაციის ფორმულა ორ მომხმარებელს შორის მსგავსების კოეფიციენტის გამოსათვლელად:

$$\text{similarity}(x, y) = \cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \times \|\vec{y}\|} = \frac{\sum_{i \in I_{xy}} r_{x,i} r_{y,i}}{\sqrt{\sum_{i \in I_x} r_{x,i}^2} \sqrt{\sum_{i \in I_y} r_{y,i}^2}}$$

შესაბამისად ვიღებთ რიცხვს 0-დან 1-მდე, სადაც 1 არის მაქსიმალური მსგავსება, ამ სკალარით შესაძლებელი ხდება რეკომენდაციების კოეფიციენტის უფრო ზუსტად გამოთვლა.

5 დასკვნა

ExploreGeorgia არის ვებ-საიტის პერსპექტიული კონცეფცია, რომელიც მიზნად ისახავს საქართველოში დაინტერესებული, ტურისტებისა და ავანტურისტების მოზიდვას და ხელშეწყობას. ამ ეტაპზე, მთავარი მოსალოდნელი პრობლემაა მომხმარებლებისა და მაშასადამე ინფორმაციის სიმწირე, თუმცა ტესტირებისთვის ეს პრობლემა გადაჭრადია ხელოვნური მონაცემებით. პროექტის წარმატების შემთხვევაში, შემდგომში გაფართოება შეიძლება, რათა სერვისი მოიცავდეს მთელ კავკასიას, ან თუნდაც გახდეს გლობალური.

წყაროები და ლიტერატურა

- [1] John S. Breese, David Heckerman, and Carl Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, 1998 Archived 19 October 2013 at the Wayback Machine W.-K. Chen, *Linear Networks and Systems*. Belmont, Calif.: Wadsworth, pp. 123-135, 1993. (Book style)
- [2] Gleick, James (2012). *The information: a history, a theory, a flood* (1st Vintage books ed., 2012 ed.). New York: Vintage Books. p. 410. ISBN 978-1-4000-9623-7. OCLC 745979816K. Elissa, "An Overview of Decision Theory," unpublished. (Unpublished manuscript)

- [3] Koren, Yehuda; Bell, Robert; Volinsky, Chris (August 2009). "Matrix Factorization Techniques for Recommender Systems". *Computer*. **42** (8): 30–37. CiteSeerX 10.1.1.147.8295. doi:10.1109/MC.2009.263. S2CID 58370896

ბიბლიოთეკები

- [1] <https://pandas.pydata.org/>
- [2] <https://numpy.org/>
- [3] <https://matplotlib.org/>
- [4] <https://scikit-learn.org/>
- [5] <https://www.tensorflow.org/>