Предлог пројекта из СИАП-а

Овај документ садржи кратак опис онога што је тема пројекта и дефиниција, мотивација за одабрану тему. Након мотивације следи преглед владајућих ставова и схватања у литератури, затим скуп података који је укратко описан. Такође је наведен и софтвер који ће бити коришћен, као и метод евалуације. На самом крају документа налази се план рада на пројекту.

Тема пројекта је систем *Capital Bikesharing* (*http://www.capitalbikeshare.com*) за изнајмљивање бицикала са станица у америчким градовима - Вашингтон, Арлингтон и Александрија, ВА и округ Монтгомерија. На располагању је преко 3000 бицикала и 350 станица. Регистровани члан изнајмљује бицикл на било којој станици и враћа га на неку од станица у систему.

Дефиниција пројекта

Предвиђање локација нових станица за бицикле и уклањање непотребних станица, анализом постојећих станица и фактора које утичу на популарност сваке појединачне станице у систему. Идеја овог пројекта је фокусирање на људима битне локације које се у одређеној мери налазе у близини станица, како би се близином неког типа локације објаснила популарност станице и пронашли шаблони који указују на то где би требало поставити нову станицу и/или објашњавају непопуларност неке постојеће станице.

Мотивација

Свака станица има одређени степен популарности који је одређен бројем изнајмљених и враћених бициклова на ту станицу. Претпоставка је да су станице које имају већи промет популарније и да доносе веће приходе компанији. За отварање и затварање било које станице, компанија мора да уложи одређена финансијска средства. Због овога је у циљу компаније да ефикасно предвиди локације на којима би могла отворити нове станице које би имале велику популарност, а да укине оне чија је популарност мала. Тиме се утиче и на финансијски аспект пословања, а и на задовољство корисника.

Преглед владајућих ставова и схватања у литератури

[1] Matthew W. C. (2014) *Predicting the Popularity of Bicycle Sharing Stations: An Accessibility-Based Approach Using Linear Regression and Random Forests*
http://www.indicatrix.org/publications/2014/Conway-Bikeshare-Accessibility.pdf

Тема рада: Формирање општег модела популарности станица над системом за изнајмљивање бициклова у Вашингтон DC региону (*Capital Bikeshare*) на основу доступности станица. Овако формиран модел користити за предикцију популарности станица у компанијама за изнајмљивање бициклова у Сан Франциску (*Bay Area Bikeshare*) и Минеаполису (*Nice Ride Minnesota*).
Подаци: Рад је у обзир узео седам параметара: радна места на 30 или 60 минута вожње од било које станице, стамбене објекте на 30 или 60 минута вожње удаљене од било које станице, радна места и стамбене објекте на 10 минута пешачења од било које станице и све станице за изнајмљивање бициклова до којих се са дате станице стиже за мање од 30 минута вожње бицикла. Подаци су изведени помоћу *OpenStreetMaps* и *OpenTripPlanner* софтвера.
Коришћени алгоритми: Линеарна регресија и Random Forest алгоритам.
Остварени резултати: Развијени модел се показао као добар за предвиђање популарности станица (како постојећих, тако и нових) у систему на којем је развијан.

Међутим, без корекције параметара модел није показао добре особине преносивости на друге системе за изнајмљивање бициклова, што је био примарни циљ овог рада.

[2] Patrick V , Jan F. E , Dirk C. M. (2014/2015) Decision support for tactical resource allocation in bike sharing systems

https://www.tu-braunschweig.de/Medien-DB/winfo/publications/decision_support_for_tactical_resource_allocation_in_bike_sharing_systems.pdf

Тема рада: Рад представља интегрисани приступ алата *Data Mining-a* и математичке оптимизације за подршку тактике за алокацију ресурса у „*Citybike Wien*“ система за изнајмљивање бицикала. Праћењем података о бициклима, о станицама на којима се изнајмљују, односно враћају, о трајању вожње током радних дана и викенда уочени су неки обрасци понашања који су искоришћени како би се бицикли равномерно распоредили по станицима, пребацивање станице, ако је потребно, уз што мањи могући трошак по компанију и уз веће задовољство корисника.
Подаци: Подаци су добијени од стране „*Citybike Wien*“ система за изнајмљивање бицикала и обухватају период од две године (2008. и 2009.година). Анализирани су подаци за летњи период (април-октобар).
Коришћени алгоритми: Алгоритам из алата *RapidMiner*-кластеровање.
Остварени резултати: Обављена студија је показала да развијени модел даје разумне нивое попуњености станица бициклима и идентификује трошкове потребне за измештање станица ако се то покаже потребним.

Скуп података

За ову тему је потребно направити скуп података над којим ће се вршити даље анализе. Предвиђено је да се скуп формира на основу података доступних на интернету. У наставку ће бити више речи о овим подацима.

Један скуп података је преузет са сајта *Capital BikeSharing* (*http://www.capitalbikeshare.com/trip-history-data*) и садржи информације о трајању путовању, датум и време почетка путовања, датум и време завршетка путовања, почетна и крајња станица, бицикл (његов ID број) и тип чланарине и то по кварталима за претходних пет година (подаци датирају од 2010. године). На основу ових података је могуће израчунати једну компоненту популарности станице засновану на промету који одабрана станица остварује током времена. То значи да би на основу ових података за сваку појединачну станицу били срачунати подаци о броју изнајмљених тј. враћених бициклова на нивоу одабране временске јединице (нпр. месец или година). Подаци су дати у csv датотеци.

Подаци о станицама садрже назив станице, географске координате (географску ширину и дужину), датум постављања станице, датум уклањања станице (само ако је станица претходно уклоњена), да ли је станица закључана и укупан број места на станици предвиђен за остављање бициклова. Подаци ће бити преузети са званичне странице компаније у облику xml датотеке (*http://www.capitalbikeshare.com/data/stations/bikeStations.xml*).

Такође је предвиђено да се у обзир узму и подаци о станицама метроа, већим компанијама и важнијим знаменитостима на територији Вашингтон DC округа. Мотивација за прикупљање података о станицама метроа лежи у томе што многи људи који користе метро као превозно средство користе и друга превозна средства како би стигли до тачки града до којих не могу стићи метроом. Међу тим превозним средствима се налазе и бициклови. Због тога би било битно прикупити податке о метро станицама у овом округу. Имена свих станица из листе станица, просечан број путника током дана и линија којој станица припада ће бити преузети са

линка https://en.wikipedia.org/wiki/List_of_Washington_Metro_stations#cite_note-ridership-8. За сваку станицу са овог линка ће се помоћу Overpass API-ја (http://wiki.openstreetmap.org/wiki/Overpass_API) пронаћи тачна локација у виду географске дужине и ширине. Подаци о компанијама су значајни на за локације станица из истих разлога као и подаци о метро станицама – људи често користе бицикле као превозно средство до свог радног места. Било би важно за све велике компаније из овог региона доћи до података о називу, локацији и типу компаније (IT, телекомуникације итд.). У овом моменту нема предлога локације на којој би се могли наћи овакви подаци. У плану је прикупљање података о другим важним локацијама са ове територије (образоовне институције, верски објекти, историјске знаменитости) како би се установио њихов утицај на положај станица за бицикле. Постоји предлог извора из којег би се могли преузети подаци, али у том случају би били доступни само подаци о важним објектима који се налазе у историјски важним грађевинама, док би остали објекти били занемарени. За сваку локацију би се бележио њен назив, тип и географске координате. У случају коришћења података из извора који ће бити наведен, до података би се дошло на идентичан начин као код података о метро станицама. Линк до извора: https://en.wikipedia.org/wiki/National_Register_of_Historic_Places_listings_in_central_Washington,_D.C. .

Софтвер

За израду пројекта ће бити коришћени *Microsoft SQL Server 2008 R2 Developer Edition* као софтвер за складиштење, обраду и анализу података. Data Mining анализе ће бити изведене у пропратним алатима Мајкрософт SQL сервера, а у случају да алат нема све потребне функционалности биће коришћен *RapidMiner* софтвер за обраду и анализу података. Апликација ће бити израђена у C# програмском језику.

Метод евалуације

Што се тиче алгоритма за анализу података, користићемо алгоритме за проналажење асоцијатвних правила и алгоритме за класификацију. Приликом иницијалне анализе података, користићемо алгоритам за проналажење асоцијативних правила да бисмо пронашли зависности у подацима. Након тога ћемо припремити податке за класификациони алгоритам да бисмо открили зависност популарности станице од места на мапи. Да бисмо евалуирали тачност наученог модела, из обучавајућег скупа ћемо изузети поједине тренутно добро оцењене станице и поједине већ затворене станице (лошег промета). Упоређивањем већ познате оцене и оцене на основу израчунатог модела можемо утврдити тачност модела. Додатно, могуће је на основу модела генерисати топлотну мапу која ће осликавати оцену потенцијалне посећености станица које би биле изграђене на одређеном месту.

План

Плана рада на овом пројекту обухвата следеће битне тачке (milestones) :

- прикупљање података, (АВ, ГГ)
- трансформација података, (АВ, ГГ)
- креирање модела, (АВ, ГГ, РТ)
- провера модела (РТ)
- визуелизација добијених резултата. (РТ)

Тим

Тим чине : **Г**орана **Г**ојић (E2 51/2015), **А**нгелина **В**уjановић (E2 21/2015) и **Р**адован **Т**уровић (RA 4 /2011).