UDC: 517.246.8 Info M str. 27-33

ПРОНАЛАЖЕЊЕ ТЕМА У РЕЦЕНЗИЈАМА ХОТЕЛСКИХ УСЛУГА
FINDING TOPICS IN REVIEWS OF HOTEL SERVICES

Нађа Булајић, QCerris
Сандро Радовановић, Универзитет у Београду – Факултет организационих наука

РЕЗИМЕ: Латентна Дирихлеова алокације (ЛДА) је један од начина статистичког моделовања тема. У оквиру ЛДА модела, документи су представљени као мешавина тема, док сваку тему чине речи са вероватноћама за сваку реч да припада датој теми. У овом раду, ЛДА ће бити представљен помоћу примера који се односи на идентификовање тема у скупу рецензија остављених од стране посетиоца хотела. Поменуте рецензије су преузете са *Tripadvisor* апликације и смештене у скуп података под називом *„Trip Advisor Hotel Review“*, где се детаљном анализом преко 20.000 рецензија откривају главне теме које најчешће карактеришу хотеле. Скуп података је процесиран и прилагођен потребама истраживања у циљу добијања бољих резултата, након чега су формирани речник података и корпус који је коришћен као улазни параметар за прављење модела. Добијени модел се састоји од листе тема и након визуелизације модела, свака од тих тема је именована. На самом крају рада, описане су ситуације у којима овај модел може наћи примену и тиме донети користи како корисницима апликације, тако и самом менаџменту хотела.

КЉУЧНЕ РЕЧИ: Моделовање тема, Латентна Дирихлеова алокација, Машинско учење

ABSTRACT: Latent Dirichlet Allocation (LDA) is one way of statistical modeling of topics. Within the LDA model, documents are represented as a mixture of topics, while each topic consists of words with probabilities for each word yes belongs to the given topic. In this paper, LDA will be introduced using an example related to topic identification in a set of reviews left by hotel visitors. The mentioned reviews are taken from the Tripadvisor application and placed in a data set called "Trip Advisor Hotel Review", where a detailed analysis of over 20,000 reviews reveal the main themes that most often characterize hotels. The data set was processed and adapted to the needs of the research living in order to obtain better results, after which a data dictionary and a corpus were formed which were used as input parameter for model building. The resulting model consists of a list of topics and after visualizing the model, each of those topics are named. At the very end of the paper, the situations in which this model can be applied and thus brought about are described benefits both the users of the application and the hotel management itself.

KEY WORDS: Topic Modeling, Latent Dirichlet Allocation, Machine Learning

УВОД

Tripadvisor је апликација која помаже њеним корисницима да се информишу о угоститељским објектима и туристичким дестинацијама. Олакшава им претрагу и потрагу идеаног места за одмор, избор ресторана у ком ће прославити важан догађај или информисање о знаменитостима које вреди посетити. За свако место или објекат који се налази у оквиру апликације, корисник може да остави коментар и пренесе своје утиске, као и да прикупи искуства других корисника апликације која ће му помоћи у доношењу одлуке. У случају да се ради о избору хотела за одмор, корисник може да се сусретне са великим бројем рецензија. Читање и анализирање сваке рецензије појединачно одузима много времена и доводи до тога да корисник након прочитаних пар рецензија одустане или због површног прегледа утисака изостави неке битне аспекте и донесе погрешну одлуку. Такође, менаџменту хотела би значило да у пар кликова може да сазна главне утиске које њихов објекат оставља на посетиоце истог.

Истраживање које се спроводи има за циљ прављење модела на основу скупа података под називом *“Trip Advisor Hotel Review*[1]*”*, који садржи преко 20.000 различитих рецензија које су оставили посетиоци хотела. Коришћењем овог скупа података је извршено моделовање тема уз помоћ латентне Дирихлеове алокације.

Моделовање тема је техника ненадгледаног машинског учења [4] која групише сличне речи и изразе који заједно могу да окарактеришу одређени скуп докумената [13]. На овај начин је могуће да се закључе теме унутар неструктуираних података. Алгоритми су задужени за стварање колекција тих речи и израза и као излаз дају листу тема, док на доносиоцу одлуке остаје да протумачи добијене теме и анализира могуће примене добијеног модела.

Латентна Дирихлеова алокација (ЛДА) [3], укратко, може да се опише као алгоритам у оквиру којег се идентификују речи које се најчешће појављују заједно у сваком документу засебно. Овим може да се дође до закључка које рецензије говоре о сличним стварима и у њима ће касније вероватно и доминирати сличне теме. Након тога се формира листа тема које описују целокупан скуп података над којим се истраживање спроводи.

За почетак, процес израде ЛДА модела је подељен на пет корака како би само настајање модела било доста разумљивије. Након што се објасни сваки од корака, приказују се добијени резултати применом сваког од њих. Добијени резултати се затим визуелно приказују како би се анализирао однос и повезаност између тема. Затим се свака од тема именује на основу контекста који заједно чине речи од којих је сачињена свака тема. На самом крају, потребно је одредити даљу примену добијеног модела у различитим ситуацијама и истакнути користи које модел пружа.

1 Скуп података можете пронаћи на следећој адреси: https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews

ПРЕГЛЕД ЛИТЕРАТУРЕ

Анализа текстуалних података са веб странице *Trip Advisor* и њему сродних веб страница је већ коришћена у практичним и научним радовима са идејом да се идентификују мишљења и преференције корисника хотелских услуга. Односно, анализом текстуалних података се долази до информације шта корисници мисле о одређеној локацији, хотелу или соби. Корисницима таква информација помаже да донесу одлуку у којем хотелу да одседну, а менаџмент хотела може да пронађе врлине и мане својих соба.

Занимљив пример примене моделовања тема јесте у раду [1] где су анализиране *Trip Advisor* рецензије лондонских музеја. Кључне идентификоване теме које су се јављале у највећем броју рецензија су именоване "деца" (заступљене кључне речи: *children; adults; interactive; year_old; activity; hands; play; age; family; loved; like*), "тешко" (заступљене кључне речи: *difficult; read; light; small; describe; display; disappoint; signs; front; reviews; board; timed; bad; frustrating; visitors*), "сати" (заступљене кључне речи: *hours; spent; least; money*) и "ред" (заступљене кључне речи: *queue; minute; wait; quickly; straight; terms; busy; arrived; year_old*). Иако за неке од ових речи можемо интуитивно да закључимо и на основу назива и кључних речи сентимент корисника, корисна информација за доносиоце одлука се добија када се укрсти тема са оценама. Тако је тема ред асоцирана са веома ниском оценом (просечна оцена испод 2,5), деца са ниским оценама корисника (просечна оцена око 2,5), док су теме тешко и сати са надпросечном оценом (око 3,5). Односно, закључак који је изведен је да корисницима лондонских хотела не смета тешка навигација кроз музеј и дугачко време проведено у музеју колико присуство деце и чекање у редовима.

Пример сличан овом раду се може наћи у раду [5] где је извршено поређење тема у коментарима са *Airbnb*-ја и *Trip Advisor*-а. Коментари су се односили на рецензије истих услуга изнајмљивања соба у Великој Британији. Теме су добијене применом латентне Дирихлеове алокације и показано је да корисници различитих веб страница имају различите коментаре. Корисници *Airbnb*-ја више пажње поклањају домаћину, транспорту до локације, чистоћи и самој локацији, док корисници *Trip Advisor*-а више пажње дају добијеној вредности, особљу, погледу и храни. У корелацији добијених тема и оценама које су корисници дали, види се да постоји разлика између корисника. Иако је очекивање да су сличне теме слично оцењене, теме попут особља имају негативну корелацију. Наиме, корисници *Trip Advisor*-а чешће дају негативне оцене због особља, док корисници *Airbnb*-ја чешће дају позитивне оцене због особља.

Поред наведена два примера, у литератури се може наћи анализа текстуалних података са циљем проналаска тема у различитим применама, попут возних станица [17], мигрантске кризе у Европи [7], тема у новинарским чланцима и слично.

МЕТОДОЛОГИЈА ИСТРАЖИВАЊА

Моделовање тема је у овом раду представљно преко латентне Дирихлеове алокације. Модел латентне Дирихлеове алокације има за циљ да пронађе групе речи које се често појављују заједно у различитим документима [11]. Након тога, од пронађених група речи формира теме, где је свака тема представљена као листа речи са вероватноћама за сваку реч да припадају датој теми. Такође, за сваки документ у скупу података је могуће одредити у ком проценту садржи коју тему, односно, која мешавина тема чини сваки документ.

Како би се процес израде самог модела боље разумео, могуће га је поделити на пет кључних корака [21, 10]. Кораци који се спроводе су следећи: прикупљање података и разумевање проблема, чишћење података и токенизација, прављење речника података и корпуса, проналажење оптималног броја тема и прављење модела и на крају, визуелизација модела и анализа добијених резултата.

У првом кораку, прикупљање података и разумевање проблема, потребно је да се направи или пронађе скуп података који одговара потребама истраживања, као и да се разуме сам проблем који се решава и постави циљ због којег се истраживање спроводи.

Након њега следи чишћење података и токенизација. Токенизација је процес раздвајања текстуалног скупа података на појединачне елементе, односно, «разбијање» сваког документа на речи од којих се састоји. Када би сваки појединачни елемент документа, без икаквог кориговања био претворен у токен, дошло би до проблема приликом тумачења резултата. Проблем настаје због тога што је за тумачење тема на прави начин потребно да теме буду сачињене од квалитетог скупа речи. Из тог разлога је потребно да урадимо чишћење и сређивање самог скупа података и на тај начин обезбедимо да се међу токенима нађу само елементи који нам касније могу бити од користи. Дакле, чишћење података подразумева уклањање свих неинформативних карактера и сређивање самог скупа података. Под неинформативним елементима, за потребе овог истраживања, могу да се сматрају сви елементи који нису речи. Дакле, потребно је отклонити електронске адресе, бројеве, знакове интерпункције и слично. Такође, како се неки токени са истим значењем написани у различитом облику не би понављали, потребно је да се сви токени пребаце у оригинални облик речи и то се врши процесом лематизације. [18, 9]

Пре прелажења на прављење речника, може да се сагледа да ли сви добијени токени заправо и потребни за добијање најбољег модела. Како неке речи не носе потребну тежину да би им се посветила пажња, било би пожељно да буду уклоњене из модела. Дакле, потербно је избацити стоп речи (енг. *Stopwords)* - речи које не носе никакве значајне информације и одређене врсте речи, односно *POS* тагове. *POS* тагови који су укључени у истраживање су именице, глаголи и придеви. Сада, након што је детаљно спроведена припрема података, може да се пређе на следећи корак. [18, 9]

Трећи корак се односи на прављење речника података и корпуса. Речник се прави од различитих токена који се налазе у скупу података над којим се спроводи истраживање.

Пролази се проз читав скуп података и сваки токен који се до тада није појавио добија своје место у речнику, као и идентификациони број који је обележава. Битан концепт који се овде појављује је филтрирање речника. Филтрирање речника подразумева уклањања елемената речника, односно речи, које се појављују у јако великом броју докумената како би се избегло да се у темама подударају најрелевантније речи, као и уклањање речи које се појављују у јако малом броју докумената. Овако добијен речник се даље користи за израду корпуса података. Корпус за сваки документ у колекцији, садржи листу токена и фреквенцију његовог појављивања у самом документу. [2, 18, 9]

Добијен корпус података представља главни елемент који је потребан за прелазак на следећи корак, проналажење оптималног броја тема и прављење модела. Како је број тема улазни параметар, да би се избегло његово насумично подешавање, потребно је да се на одређени начин пронађе његова оптимална вредност. У овом раду, оптималан број тема је пронађен комбинацијом мере кохерентности ‹*c_v*’ и Џакардове сличности. Како мера кохерентности одређује степен сличности између речи унутар тема, који је пожељено да биде што већи, а Џакардова сличност одређује степен сличности између тема, за коју је пожељно да има што нижу вредност како би теме покриле што већи део скупа података, за оптималан број тема је најбоље прогласити онај број тема који има највећу разлику између поменутих мера. Затим, уз пронађен оптималан број тема и подешавањем осталих параметара може се приступити креирању самог модела. [2, 18, 9]

У петом кораку је потребно да се визуелним путем, прикаже однос између добијених тема, а након тога уз помоћ листе најрелевантнијих речи, свакој теми додели назив и да се изврши дискусија о добијеним резултатима.

Целокупно истраживање, спроведено кроз претходно описане кораке је урађено у оквиру *Jupyter Notebook* окружења, коришћењем *Python* програмског језика. Библиотеке које су коришћене у раду су следеће:

- *pandas* [12] - за учитавање скупа података и манипулацију са подацима;
- *spaCy* [19, 16] - за припрему скупа података (такође је коришћен и њен уграђен модел *en_core_web_md*, обучен за рад са текстуалним подацима на енглеском језику;
- *NLTK* [14, 6]- за коришћење *POS* тагова, токенизацију, лематизацију и стоп речи
- *gensim* [15, 16] - за израду речника, његово филтрирање, прављење корпуса, рачунање мере кохерентности, израду самог модела (*LdaMulticore*) и визуелизацију самог модела (*pyLDAvis*).

РЕЗУЛТАТИ ИСТРАЖИВАЊА И ДИСКУСИЈА РЕЗУЛТАТА

У складу са описаним корацима у претходном поглављу, у наставку ће бити приказани добијени резултати у оквиру спроведеног истраживања.

Скуп података над којим је спроведено истраживање је «Trip Advisor Hotel Review”. У поменутом скупу података налази се 20.491 опсервација од којих свака описује искуство одређеног корисника апликације о времену проведеном у неком од хотела које апликација покрива.

Након што је скуп података учитан и изанализиран, приступљено је припреми података за прављење корпуса, који касније служи као улаз у модел. Припрема обухвата следеће ставке:

1) провера да ли свака опсервација у скупу података садржи рецензију, односно да ли нека опсервација може да се означи са недостајућом вредношћу
2) пребацивање свих речи у речи написане малим словима
3) елиминисање свих елемената који нису речи из скупа података

Сада, када је скуп података «растерећен» уклањањем непотребних елемената, урађена је токенизација. Иако се у скупу података тренутно налазе само речи, то не значи да су све оне довољно релевантне како би се њима дефинисале теме. Како би се избегло појављивање речи које не носе довољно информација или понављање истих речи више пута написаних у различитом облику, пожељено је да добијени токени прођу кроз још једно сређивање пре него што се искористе за прављење речника.

Дакле, након токенизације скупа података, уклоњене су стоп речи, урађена је лематизација и остављени су само POS тагови који карактеришу именице, глаголе и придеве. Скуп података је сада спреман за прављење речника, који ће уз још пар ситних корекција бити добра основа за прављење корпуса података.

Почетник речник садржи 33.511 елемената. Добијени речник је превелик за коришћење. У случају да га користимо у целости за прављење корпуса, већ сада можемо закључити да теме добијене на овај начин неће носити много корисних информација. Разлог томе је што се велики број речи појављује јако често и у готово свим рецензијама, самим тим, може да се закључи да ће се одређене речи налазити у готово свим темама, што ће знатно отежати прављење разлика међу темама, а касније и именовање тема. Како би се ово спречило пожељно је да се уради филтрирање речника којим ће бити уклоњене описане речи. Филтрирање је прво рађено постављањем горње границе тако да речник обухвата само оне речи које се појављују у мање од 80% рецензија. У другој итерацији овај проценат је смањен на 50% и како се тиме губи само неколико додатних речи, ова вредност је задржана до краја истраживања.

Постављањем горње границе је решен проблем превише учесталих елемената речника, али не смеју да се занемаре и елементи који или немају никакав или имају јако мали значај за само истраживање. Поменуте речи неће бити препознате као релевантне и изабране за доминантне речи приликом дефинисања тема али како постоји јако велики број речи које се налазе у само једној или свега неколико рецензија, могу знатно да успоре прављење модела. Изостављање оваквних елемената из речника је постигну-

то постављањем доње границе на 200, односно, све речи које се појављују у мање од 200 рецензија су избачене из речника. Додатна предност постављања доње границе се огледа у томе да ће одређена реч која је погрешно написана и случајно залутала и постала елемент речника – бити уклоњена из истог.

Још једна измена која је направљена је ограничавање речника да садржи само елементе који садрже више од два карактера.

Последњи корак у сређивању речника представља пролажење кроз цео речник како би се уочило да ли се у њему и даље налазе одређени елементи који могу да отежају добијање што бољег модела. У овом проласку је уклоњено 40 додатних речи које су наведене у наставку уз навођење разлога за њихово уклањање.

- *'review', 'reviewer'* – изостављају се из анализе јер је јасно да се у случају када се анализа спроводи над скупом података о рецензијама, заправо и говори о рецензијама и није потребно да се то додатно напомиње у темама које ће касније бити генерисане;
- *'minute', 'min', 'pm', 'hour', 'night', 'day'* – како су у истраживање укључене само речи, а временске одреднице најчешће иду из одређени број (број дана проведених у објекту, број сати или минута чекања нечега, време када се десио одређени догађај...), ови појмови не могу да буду од користи у случају да се нађу у оквиру неке од тема;
- *'location', 'place'* – свака рецензија садржи утисак о месту које су корисници апликације посетили; самим тим, спомињање ове две наведене речи не доноси корисне информације;
- *'star'* – наведена реч говори о квалитету хотела израженом бројевима од један до пет; како су бројеви изостављени из истраживања, поменута реч сама не може да допринесе даљој анализи;
- *'stay', 'give', 'pleased', 'say', 'show', 'none', 'part', 'make', 'maker', 'take', 'get', 'tell', 'ask', 'include' 'know', 'use', 'self', 'think', 'thank', 'stayed', 'need', 'try', 'thing', 'nothing', 'add', 'non', 'left', 'send', 'time'* – додатне речи, већином глаголи који се јако често употребљавају али не садрже корисне информације за проблем који се решава у оквиру овог истраживања.

Речник припремљен на овакав начин сада, уместо преко 33.000 елемената различитог степена релевантности, сада обухвата 996 елемената. Како је сређивању речника приступљено са посебном пажњом, остале су само најрелевантније речи које би требало да знатно олакшају процес разумевања тема и даљи сад са њима.

Израда корпуса не захтева посебну припрему и подешавања параметара већ само речник као улазни елемент и након кратког задржавања на овом кораку може да се приступи изради самог модела.

Поред корпуса, јако битан улазни елемент алгоритма је број тема. Број тема је у овом раду одређен коришћењем мере кохерентности *'c_v'* и Џакардове сличности. Оптималан број тема је биран у опсегу од 1 до 30. Подешавањем осталих параметара на различите начине, добијена су два модела која се могу узети у обзир за прављење оптималног модела. Први модел садржи 25 тема, док други садржи 28 тема. Како би се донела одлука који од ова два модела ће се користити у даљем истраживању, у наставку ће табеларно бити приказане разлике у мерама сличности између ова два модела. Мера сличности која је такође коришћена, а до сада није помињана је перплексност (енг. *perplexity)*. Перплексност представља количину збуњености – што је збуњеност већа, то су резултати изненађујући, тако да је пожељно да ова вредност буде што мања.

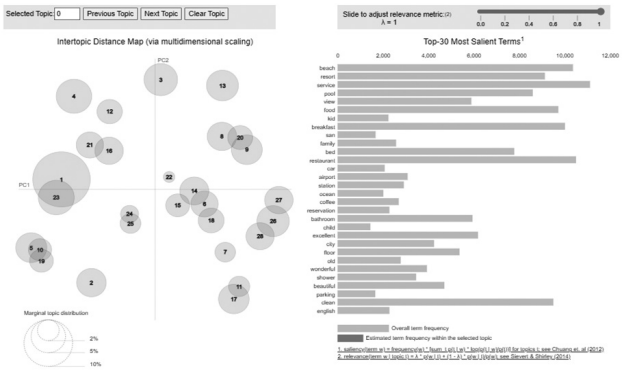
Табела 1: Поређење два модела моделовања тема

Модел	'c_v'	перплексност	Џакардова сличност
Модел са 25 тема	0,4321	-6,3573	0,0445
Модел са 28 тема	0,4569	-6,3456	0,0486

Из табеле 1 се види да нема великих варијација између посматраних модела када се упореде мере сличности. Ниједан модел није доминантан у све три мере, нити садржи велику предност у некој од мера појединачно. Коначна одлука је из ових разлога донесена упоређивањем визуелизација поменутих модела. Модел са 28 тема има мање преклапања и боље распоређене теме. Иако ове разлике нису драстичне и вероватно не би била грешка користити и модел са 25 тема, даље истраживање је настављено избором модела са 28 тема за оптималан модел.

Уз одређивање оптималног модела, остао је још један корак – именовање добијених тема и сагледавање даље примене модела.

Именовање тема је започето анализом тема које имају преклапања. Именоване су по групама од две, три или четири теме. Када се именују теме које се преклапају потребно је водити рачуна да се избегава коришћење само речи које су доминантне у свим темама како би се спречило да више тема носи иста или слична имена јер тиме смањујемо њихову корисности код примене модела. Из тог разлога је пожељно да свака тема садржи неко обележје које је карактеристично само за њу. То обележје може да буде реч која има се у највећој мери налази само у једној теми или реч која означава контекст који дели више речи у оквиру теме.



Слика 1. Моделоване теме и заступљеност речи у њима

Као пример за горенаведено може да се узме реч *'city'*. У већем броју тема ова реч се налази у листи најрелевантнијих. Када би се користила приликом именовања, као резултат би био добијен низ тема са истим именом што доводи до лошег коришћења модела у пракси чак иако је модел сам по себи добар. Уместо овога, ако се за називе употребе имена конктретних туристичник локације, добијене теме ће дати значајније корисније информације. Још један корак који може да се предузме да називи тема буду још одређенији јесте да се уз туристичку дестинацију дода још неки опис у називу теме. Овим се доприноси томе да у случају да одређена рецензија у неком проценту садржи тему која у називу има конкретно место, а објекат који се описује се не налази у њему, ипак доси корисну информацију. Конкретни примери за именовање оваквих тема су: *„Florence & Comfortable"*, *„Punta Cana & Bad hygienic"*, *„Quite area & Boston"* и слично. Графички приказ добијених тема се налази на слици 1. Приказане су теме пројектоване на два главне компоненте (добијене анализом главних компоненти), где величина теме представља њену заступљеност у подацима. Са десне стране слике су приказане најзначајније речи.

Именовање тема које су појединачне је рађено комбиновањем речи које се у највећем проценту налазе само у тој теми и контекста у којем могу да се нађу остале речи. Пример именовања једне од таквих тема је тема број 10 у којој се у највећој мери појављују следеће речи: *'kid'*, *'adult'*, *'family'*, *'child'*, *'daughter'*, *'son'*. Наведене речи недвосмислено говоре о породици, самим тим је и тема добила назив *„Family vacation"*.

Детаљном анализом осталих тема и листа речи које их карактеришу, на сличан начин су именоване и све преостале теме, чији се називи могу наћи у табели испод.

Табела 2: Идентификоване теме у рецензијама

Тема	Назив теме	Тема	Назив теме
Topic 1	„Resort & Beach (Pool) & Meal"	*Topic 15*	„Birthday celebration"
Topic 2	„Wonderful & Honeymoon"	*Topic 16*	„Business trip"
Topic 3	„Dirty rooms"	*Topic 17*	„Florence & Comfortable"
Topic 4	„Rude staff"	*Topic 18*	„Public transport & Paris"
Topic 5	„Punta Cana & Bad hygienic"	*Topic 19*	„Wedding & Couple location"
Topic 6	„Harbour, Hong Kong, Sydney & Singapore"	*Topic 20*	„Breakfast & Buffet table"
Topic 7	„San Juan, San Francisco & Puerto Rico"	*Topic 21*	„Bad service"
Topic 8	„Parking & Smoke"	*Topic 22*	„Palace & Tokyo"
Topic 9	„Apartment"	*Topic 23*	„Resort, Beach (Pool) & Excellence"
Topic 10	„Family vacation"	*Topic 24*	„Beautiful balcony view, Romance & Expedia"
Topic 11	„Suitable environment"	*Topic 25*	„Beautyful balcony view & Waikiki"
Topic 12	„Relaxing weekend"	*Topic 26*	„Urban location, Amsterdam, London & Berlin"
Topic 13	„Bathroom & Room"	*Topic 27*	„Quiet area & Boston"
Topic 14	„Villa & Bali"	*Topic 28*	„Turist area & Spain"

Како је већ напоменуто, свака тема се састоји од мешавине различитих тема са вероватноћом да одређена рецензија припада свакој од њих. У наставку је приказан пример две рецензије и тема од којих се састоје како би ово било јасније.

Табела 3: Примери рецензија и добијених тема

Рецензија	Вероватноћа припадности темама
Пример 1: ‹nice hotel expensive parking got good deal stay hotel anniversary, arrived late evening took advice previous reviews did valet parking, check quick easy, little disappointed non-existent view room room clean nice size, bed comfortable woke stiff neck high pillows, not soundproof like heard music room night morning loud bangs doors opening closing hear people talking hallway, maybe just noisy neighbors, aveda bath products nice, did not goldfish stay nice touch taken advantage staying longer, location great walking distance shopping, overall nice experience having pay 40 parking night, ‘	8 - 0.675: „Parking & Smoke" 13 - 0.225: „Bathroom & Room" 24 - 0.078: „Beautiful balcony view & Romance & Expedia"
Пример 2: ‹shame hotel wasnt good restaurant, arrived clift late afternoon struggle luggage 3 bags, reception staff unhelpful uninterested, eventually managed sorted shown room 9th floor, room suite tried make separate living room putting curtain inbetween bedroom living room, bathroom tiny dirty, stayed mum unfortunatley night didnt feel suffering bad foot, decided phone reception ask doctor come hotel told ther wasnt local receptionist closest told phone, eventually decided hospital just safe, came hospital evening doormen talking girls outside let, following night ate hotel restaurant aisa cuba fantastic, think hotel intrest restaurant bar, end day sleeping ignored wouldnt stay, ‘	4 - 0.318: „Rude staff" 3 - 0.204: „Dirty rooms" 9 - 0.176: „Apartment" 6 - 0.165: „Harbour, Hong Kong, Sydney & Singapore" 21 - 0.115: „Bad service"

Уз помоћ првог примера из горење табеле, могуће је сагледати предност тога што је свака рецензија представљена кроз мешавину различитих тема. Када би рецензије била само тема број 9 - "*Apartment*", то вероватно не би било довољно да се донесе закључак о искуству особе која је оставила рецензију. Међутим, када се наведеној теми придруже теме "*Rude staff*", "*Dirty rooms*", "*Bad service*", јасно је да се у конкретном случају говори о незадовољству о посећеном апартману.

Последња и најбитнија ставка се односи на примену самог модела. О овом делу је пожељно размишљати пре него што се уопште приступи изради модела како би се на самом почетку знало шта је крајњи циљ истраживања и у ком смеру га је потребно усмерити.

Примена ЛДА модела може да се посматра из две перспективе. Прва се односи на корист коју модел доноси корисницима *Tripadvisor* апликације, док друга говори о користи коју има менаџменту хотела. Бенефит који је кључан и који доноси корист обема страна се односи на уштеду времена. Како ЛДА модел генерише теме за сваку рецензију, тиме ствара могућност да се открије о чему говори нека рецензија без потребе да се иста прочита.

Корисници апликације доносе одлуку о хотелу у ком ће одсести и употреба направљеног модела може да им олакша и као што је већ напоменуто, скрати време поменутог

процеса. Дакле, када корисник бира објекат у којем жели да борави може да се ослони на утиске које су посетиоци самог објекта оставили о времену проведеном у њему. У случају да објекат има свега неколико остављених рецензија, читање сваке од њих не представља проблем. Проблем настаја када корисник жели да прикупи што више информација али како би то урадио мора да прочита више хиљада рецензија што изискује јако много издвојеног времена. Помоћу направљеног модела, уместо да корисник чита све рецензије редом довољно је да за рецензије погледа од којих су тема сачињене и да добије потребне информације. Такође, у случају да не жели да пролази кроз сваку рецензију посебно и читањем листе тема од којих се састоји свака од њих ствара слику о хотелу, постоји могућност да изабере жељени хотел и без залажења у низ рецензија погледа листу најзаступљених тема о којима говоре остављене рецензије и тиме, у свега неколико десетина секунди, сазна шта остали корисници мисле о поменутом хотелу.

Још једна могућност која олакшава потрагу за идеалним хотелом јесте уколико корисник већ има идеју какав хотел тражи. Избором жељене теме из листе могућих тема, врши се филтрирање и корисник као резултат добија предлоге тема које га занимају.

Други начин за коришћење модела се односи на корист коју сам хотел добија. Прегледом тема које се појављују о рецензијама и рангирањем тих тема по учесталости појављивања, на једноставнији и бржи начин менаџмент хотела може да добије јасну слику о стварима које су оставиле позитивне или негативне утиске на госте који су боравили у хотелу. На овај начин могу да изаберу теме које говоре о негативним искуствима и тако дођу до одређених рецензија чијом анализом откривају шта је то што гостима смета у вези са хотелом како би лоше ствари могли да поправе и тиме унапреде пословање објекта којим управљају.

Са друге стране, у случају да су теме које карактеришу одређене рецензије позитивне, то може да доведе до закључка да, ако је одређени гост хотела приметио и ценио то довољно да о томе остави рецензију, вероватно ће и остале кориснике апликације помињање тих ствари привући ка одређеном објекту. Из тог разлога, не би било лоше да се рецензије које се састоје од похвала упућених хотелу или одређени елементи тих рецензија издвоје и користе у маркетиншке сврхе како би се скренула пажња на хотел и како би хотел привукао још гостију.

ЗАКЉУЧАК

Главни циљ рада јесте упознавање са једним сегментом моделовања тема под називом латентна Дирихлеова алокација, са стављањем акцента на процес изградње самог модела, а касније, након именовања свих тема добијених као излаз из направљеног модела и сагледавање различитих начина његове примене.

Након направљеног увода у рад, описани су сви кораци које је потребно спровести како би се направио модел. За почетак је потребно упознавање са скупом података и разумевање проблема. Скуп података који је коришћен у раду је *"Trip Advisor Hotel Review"* и садржи преко 20.000 опсервација над којим је спроведено истраживање. Следећи корак се односи на припрему података која је обухватила сређивање података за процес токенизације који је затим послужио за добијање речника. Затим је речник филтриран и ослобођен још неких додатних речи које се нису показале као корисне за процес креирање тема у оквиру истраживања. Направљени речник се даље користио за прављење корпуса који је улаз у ЛДА модел.

Током израде модела, испробана су различита подешавања параметара, као и различити начини за проналажење оптималног броја тема. Изграђен оптимални модел садржи 28 различитих тема, од којих је свака представљена кроз листу речи са вероватноћама за сваку од њих да припадају одређеној теми. Након што је добијена листа од 28 тема као излаз из модела, свака од добијених тема је именована како би модел имао сврху у даљој примени. Именовање тема је извршено на основу сагледавања речи које се налазе у оквиру сваке теме, а затим тражења одговарајућег контекста за поменуте речи.

Као закључак и суштина спреденог истраживања, на самом крају рада описани су начини примене модела од стране корисника апликације и менаџмента хотела. Такође, као и код сваког модела и у овом случају је могуће унапредити модел, испробавањем неких нових концепата, «појачаним» процесом припреме података или детаљнијом анализом направљеног речника.

ЛИТЕРАТУРА

[1] Alexander, V. D., Blank, G., & Hale, S. A. (2018). TripAdvisor reviews of London museums: A new approach to understanding visitors. Museum International, 70(1-2), 154-165.

[2] Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. John Wiley & Sons.

[3] Blei, D., Ng, A., & Jordan M. (2003). *Latent Dirichlet Allocation*, the Journal of machine Learning research 3, str. 993–1022.

[4] Delibašić, B., Suknović, M., & Jovanović, M. (2009). *Algoritmi mašinskog učenja za otkrivanje zakonitosti u podacima*. Fakultet organizacionih nauka, *Beograd*.

[5] Gao, B., Zhu, M., Liu, S., & Jiang, M. (2022). Different voices between Airbnb and hotel customers: An integrated analysis of online reviews using structural topic model. *Journal of Hospitality and Tourism Management*, *51*, 119-131.

[6] Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). *Natural language processing: python and NLTK*. Packt Publishing Ltd.

[7] Heidenreich, T., Lind, F., Eberl, J. M., & Boomgaarden, H. G. (2019). Media framing dynamics of the 'European refugee crisis': A comparative topic modelling approach. *Journal of Refugee Studies*, *32*(Special_Issue_1), i172-i182.

[8] Jacobi, C., Van Atteveldt, W., & Welbers, K. (2018). Quantitative analysis of large amounts of journalistic texts using topic modelling. In *Rethinking Research Methods in an Age of Digital Journalism* (pp. 89-106). Routledge.

[9] Jo, T. (2019). Text mining. *Studies in Big Data*. Springer, Verlag.

[10] Kherwa, P., & Bansal, P. (2019). Topic modeling: a comprehensive review. *EAI Endorsed transactions on scalable information systems*, *7*(24).

[11] Malik, U., Goldwasser, M., & Johnston, B. (2020). *Python mašinsko učenje.* Kompjuter Biblioteka, *Beograd.*

[12] McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, *14*(9), 1-9.

[13] Pascual, F. (2019) Topic Modeling: An Introduction [https://monkeylearn.com/blog/introduction-to-topic-modeling/, датум приступа: 12.10.2022.]

[14] Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook* (Vol. 9). Birmingham: PACKT publishing.

[15] Řehůřek, R., & Sojka, P. (2011). Gensim—statistical semantics in python. *Retrieved from gensim.org*.

[16] Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.

[17] Taecharungroj, V. (2022). An analysis of tripadvisor reviews of 127 urban rail transit networks worldwide. *Travel Behaviour and Society*, *26*, 193-205.

[18] Talib, R., Hanif, M. K., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11).

[19] Vasiliev, Y. (2020). *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

[20] Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

[21] Wallach, H. M. (2006, June). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 977-984).

Нађа Булајић, QCerris
Контакт: bulajicn98@gmail.com
Област интересовања: Машинско учење, Инжењерство података, Откривање законитости у подацима

др Сандро Радовановић, доцент, Универзитет у Београду – Факултет организационих наука, Јове Илића 154, Београд
Контакт: sandro.radovanovic@fon.bg.ac.rs
Област интересовања: Машинско учење, Откривање законитости у подацима, Системи за подршку одлучивању, Теорија одлучивања