Институт за математику и информатику
Природно-математички факултет
Универзитет у Крагујевцу

Студијски истраживачки рад

Примена алгоритма моделовања тема из текстуалних садржаја за проналажења одговора на постављено питање

Студент:
Јелица Васиљевић

Професор:
др Милош Ивановић

Август 2015.

Садржај

Глава 1

Увод

1.1 Опис проблема

Проналажење одговора на постављено питање претстваља свакодневни проблем. Сваким Гугл упитом, покреће се низ алгоритама коју покушавају да одгонетну шта упит заправо представља , које би странице биле релевантне и у ком редоследу. Мера "доброг"одговора на постављено питање знатно може да варира у зависности од сврхе система. Интуитивно, тематика одговра и питања може да послужи као добар критеријум одабира квалитетних одговра. Задатак овог рада је испитивање да ли заиста тематика може да помогне у проналажењу адекватног одговра и у колико мери.

1.2 Опис решења

Циљ овог рада је израда прототипа програма који би коришћењем алгоритама за моделовање тема из базе потенцијалних одговора проналазио најбољи одговор за задато питање. Дакле, програм не треба да "осмисли"одговор на задато питање већ само да "препозна"који од могућих одговора највише одговара постављеном питању

Примена оваквог решења могла би да буде значајна у различитим областима од комерцијалних до научних. На пример, омогућило би се ефикасно аутоматско одговарање на често постављена питања која могу имати различиту формулацију или ефикасно проналажење адекватних научних радова.

У раду су обрађивани текстови на енглеском језику али због природе модела, развијени програм и резултате могуће је применити и на било који други језик. Поред основног текста питања и одговора, у раду је испитан и утицај додавања синонима на проналажење одговора као и утицај свођења речи на коренске (склањање глаголских и именских наставака , енг. *stemming*)

Као компаративнии модел коришћен је приступ проналажења одговора на основу броја заједничких речи (енг. *WordCount*)

1.3 Терминологија

Општи преглед заначења термина који су коришћени у раду као што су реч, речник, вокабулар, корипус итд.

- Тема : скуп речи које је најбоље карактеришу. На пример тема рачунарсво би представљала скуп речи : алгоритам, процесор, кодирање, израчунавање, меморија, рачунар, бит, бајт, лаптоп итд. Важно је приметити да неке речи могу припадати у више тема, као што је нпр. реч израчунавање која може припадати и области математика.

- Речник или корпус - скуп свих различитих речи које се јављају у неком скупу докумената
- ....

Глава 2

Алгоритам моделовања тема

Сваким даном повећава се количина доступних дигиталних информација. Парадокс данашњег времена је да се упрокс великој количини података из различитих области све теже долази до података који су од интереса. Дакле, потребно је пронаћи алат којим би се велике количине података организовале а самим тим боље разумеле и лакше претреживале.

Тренутно, најпопуларнији начин претреге је према кључним речима. Кључне речи се предају неком систему за претрагу а као резултат добијамо скуп докумената који су повезани са њима. Иако овакав систем ради јако добро и са великом поузданошћу, постоје и другачији приступи

Најчешће питање које се поставља захтева одговр из неколико, ужих или ширих области. Међутим, кључне речи које се предају као критеријум претраге могу карактерисати и области које нису од интереса. На пример, Гугл претрга за кључне речи "ген, еволуција"ће у највећем броју случајема садржати веб странице посвећене биологији или сродним областима. Међутим, поменуте речи припадају и области рачунарства (генетски алгоритми) али ти резултати ће знатно слабије бити заступљени у односу на резултате везане за биологију. Тренутно се оваква врста проблема може решити додавањем још неке кључне речи која припада захтеваној области (на пример за задате речи : алгоритам, кодирање, програм итд.) међутим постваља се питање избора адекватних додатних речи, њихове заступљености у захтеваној области као и релевантности у односу на друге кључне речи (на примеру гугл претраге, често се дешава да механизам за претраживање испусти неку од речи и прикаже само резулатат за остатак упита)

Један од начина да се поменути проблем реши је и претрага по тематици или теми докумената. Дакле, уместо претраге по кључним речима, најпре се пронађе област у којој се врши претраживање а затим се претражују документи који припадају тој области. На тај начин елиминишу се документи који би се у резултату појавили на основу предатих кључних речи али који нису повезани са тематиком у којој се тражи одговор.

На пример, нека је потребно пронаћи све чланке листа "Политика"у којима се говори о спортским успесима наше кошаркашке репрезантиције.Сви чланци овог листа могу се поделити у неколико категорија : политика, хроника, спорт, време итд. Пошто је од интереса категорија спорта, у обзир претраге долазе једино спортски чланци са тематиком кошарке. На овај начин могуће је пратити како се мењао успех кошаркашке репрезантиције са временом, колико се обаћала пажња на такве догађаје у ком временском периоду итд.

Описани вид претраге подразумевао би да се сваки лист "Политике"најпре прочита а затим раздвоји по категоријама и "разуме"шта која категорија представља како би се добили тражени подаци. Због велике количине података, овако нешто је немогуће урадити без помоћи рачунара. Алгоритми моделовања тема представљају први корак ка решавању оваквих и сродних проблема.

Циљ алгоритама за моделовање тема је "отривање"тема присутних у некој колекцији докумената. У суштини, алгоритми за моделовање су статистичке методе које на основу

анализе свих речи у документу откривају које су то теме заступљене у том документу. За рад ове методе није потребно никакво претходно означавање докумената, теме докумената зависе једино од речи које се јављају у тексту.

Моделовање тема омогућава организацију велике количине података на нивоу који је тешко дохватљив људским могућностима.

Надаље ће се паралелно употребљавати две еквивалентне ознаке - моделовање тема или TM (скраћеница од енг. *topic model*)

2.1 *Latent Dirichlet allocation*

Latent Dirichlet allocation, надаље LDA, је најједноставнији приступ проблему моделовања тема [?] а његова примена била је и предмет овог рада.

Добро је познат роман Бранка Ћопића "Орлови рано лете". Уколико би неко ко није прочитао ови књигу желео да зна "о чему се ради"у њој, највероватније би добио одговор да је у питању књига која се бави доживљајима групе дечака на почетку Другог светског рата. Иако је то најшири оквир романа,у њему су присутне и теме о љубави, дружењу, пријатељству, рату, пустоловинама итд. Дакле, роман ,опште гледано, обухвата више тема, али се са неколико њих интензивно бави.

Уколико сада посматрамо овај рад као један пример документа, у њему се највише "говори"о моделовању тема и њиховој примени, али исто тако, само у мањој мери, и о књижњвности (пример књиге "Орлови рано лете "). Дакле, тешко је пронаћи било какав документ који се бави само једном темом. Чак се и у радовима који се баве неким уским научним областима, могу пронаћи делови који се могу сврстати и у неке друге научне области, било из исте, сродне или потпуно различите научне гране.

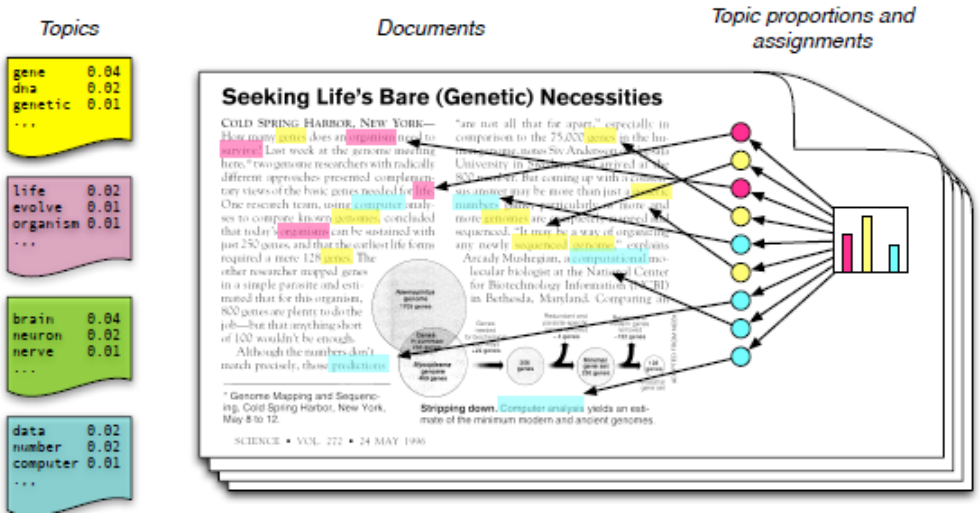
Основна претпоставка LDA модела је управо да се сваки документ може сврстати у више области, тј. да се бави са неколико тема.

Поставља се питање како то, да након читања неког документа, знамо да су у документу присутне различите теме и које су то теме. Уколико неки рад који се бави тематиком машинском учења прочита човек који се не бави рачунарством, највероватније је да би он као теме рада издвојио рачунарство и математику. Међутим, ако би исти рад прочитаоо неко ко се бави рачунарством, тада би он могао као потенцијалне теме да изведе једну или више области машинског учења, алгоритме, базе података, статистуку, теорију бројева итд. Дакле, издвојене теме зависе од "знања"онога ко чита рад. Уопштено, у оба случаја теме су издвојена на основу речи које се најчешчће срећу у областима које појединац познаје. Дакле, може се уопштити да тема представља **скуп карактеристичних речи.**.

На следећој слици представљен је чланак "*Seeking Life's Bare(Genetic) Necessities"* који "говори о"употреби анализе података за одређивање броја гена који организам треба да има да би преживео (у еволутином смислу).Може се уочити да су три најзаступљеније области у овом тексту - анализа података, еволутивна биологија и генетика. На њој су "ручно"означене неке речи које припдају овим областима. Речи које се могу сврстати у област *анализе података* су означене плавом бојом, речи које припадају *генетици* су означене жутом бојом док су речи које се односе на *еволутивну биологију* означене розом бојом. Уколико би се ова процедура применила на сваку реч текста, јасно би се уочило колико је која тема заступљена у овом тексту. Математички, "*присуство*"теме у тексту се означава односом броја "обојених"речи у једну боју и укупног броја речи у тексту.

Наравно, постоје речи које се могу сврстати у више од једне теме. Такве речи ће бити обојене са две или више боја,али због прегледности слике, такви случајеви су изостављени.

На левој стани слике дате су неке теме (енг. *Topics*) које са одређеним вероватноћама садрже неке речи. На основу припадности речи темама, извршава се описани процес означавања ("бојења") речи да би се на крају добили удели тема у тексту (десна страна



Слика 2.1: Пример чланка, преузето са [?]

слике, енг. *Topics*) Важно је приметити да у тексту постоји доста речи које не одређују ни једну конкретну тему и које се са готово истим, малим вероватноћама могу сврстати у сваку од њих. Такве речи су нпр. везници, личне заменице ,прилошке одредбе итд. Оне се једним именом називају енг. *stop words*. Како присуство таквих речи не утиче на тематику документа, то их при математичкој анализи текста треба занемарити.

LDA је статистички модел који формално описује описани процес означавања речи. Да би се у потпуности разумело како LDA ради, потребно је упознати се са *генеративним процесом* - процесом којим се креирају документи са становишта LDA-а.

2.1.1 Генеративни процес

Нека је дат неки скуп речи најчешће коришћених у научним областима математике, физике, хемије, музике, рачунарства и биологије. Нека је, даље, потребно креирати документ који има одређен број речи из датог скупа речи али тако да он највише "говори о"математици и музици, али поред ових тема, у мањој мери, "говори о"физици и програмирању, док се остале теме занемарљиво мало помињу. Како је дат фиксан скуп речи - речник (вокабулар), могуће је свакој речи придружити *вероватноћу* припадања свакој од тема. Тако ће на пример реч *интеграл* имати велике вероватноће припадања темама математика и физика, мање вероватноће у темама хемија и рачунарство док ће се са јако малим вероватноћама јављати у осталим темама. Дакле, једна реч припада свакој од датих тема али са различитим вероватноћама.

Генерисање траженог документа могуће је извести следећом процедуром :

1. Одабрати расподелу тема у документу - прецизирати која тема се са којим уделом појављује у документу. У конкретном примеру, расподела би могла бити : математика 30%, музика 30%, физика 20%,програмирање 15%, биологија 2% и хемија 3%.

2. Докле год није достигнут тражени број речи

   (a) Изабрати тему из дистрибуције која је одабрана у 1.

   (b) Изабрати реч из теме која је одабрана у 2.a). Како свака тема има речи које фаворизује (веће вероватноће у односу на остале речи), то ће се те речи највероватније одабрати у овом кораку.

Описана процедура може се графички илустровати претходном сликом (5.7). Одабрана расподела тема у документу (тачка 1 описане процедуре) предсављена је хистограмом на десној страни слике. Обојени кругови представљају одабир теме из документа (корак 2.а)) док речи повезане стралицама са њима представљају одабирану реч из те теме (корак 2.б)).[1]

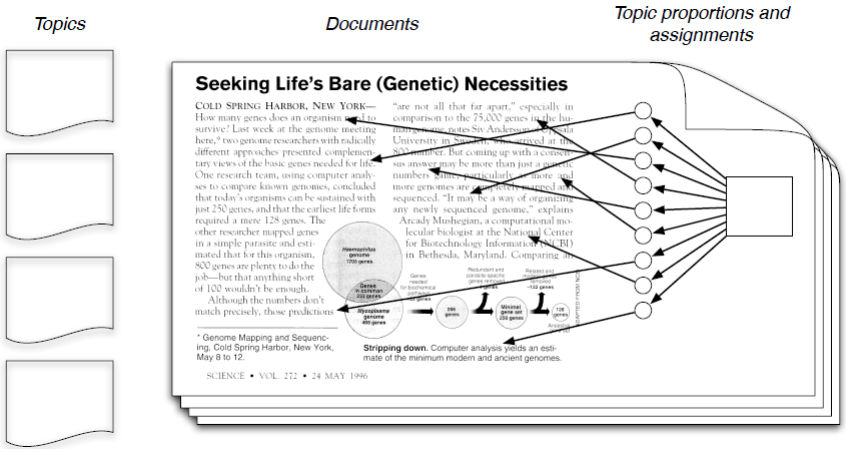
Формално, тема се дефинише као расподела речи над неким фиксним скупом речи - речником. Рецимо, тема биологија ће са већом вероватноћом садржати речи везане за ту област нпр. биљка, животиња, ћелија, ген итд. док ће тема везана за математику ове речи садржати са нижим вероватноћама у односу на речи нпр. број, разломак, променљива,коцка итд.

Дакле, основна карактеристика LDA-а је то што сви документи деле **исти скуп тема** али сваки документ те теме садржи у различитим односима. Овакаво посматрање докумената јако је природно и интуитивно.

2.1.2 Како се откривају теме - упрошћен пример

Проликом генерисања поменутог документа, било је познато која тема садржи које речи као и у којим односима је заступљена свака тема у тексту. Циљ алгоритама за моделовање тема је да **аутоматски** "открије"које су то теме присутне у неком документу и које речи припадају којој теми али **само** на основу речи које се јављају у документу, без било каквог додатног знања. Сазнање о томе која тема се у којој мери налази у неком документу, није од превеликог практичног значаја. Међутим, уколико је на располагању огромна количина докумената (нпр. дигитална база свих издаља листа "Политика"), откривање сродних докумената, или докумената који се баве само одређеним темама може бити јако важно. Због тога ће се надаље говорити о скупу докумената над којим се извршава моделовање тема, уместо о једном документу. Притом, наравно, и даље важи претпоставка да сваки документ "говори о"свим темама које се могу издвојити из свих докумената, само у различитим односима.

Према свему реченом, једино што је *видљиво*, енг. ***observed*** су документи, односно речи које се у документима јављају. Тематска расподела по документима, као и расподела речи по темама су ***скривене или невидљиве***, енг.*non observed,hidden* (Слика 2.2)



Слика 2.2: Пример чланка, преузето са [?]

Основни задатак алгоритма је окривање скривених структура на основу видљивих. У овом тренутку, моделовање тема можемо посматрати као **обрнути** генеративни процес.

[1]Расподела на основу које се одабирају пропорције тема у кораку 1, назива се Дирихлеова раподела, енг. Dirichlet distribution. На основу те одабране расподеле, врши се придруживање речи документима, енг. allocate

Дакле, циљ моделовања тема је откривање скривених структура из којих су **највероватније**, генеративним процесом, добијени видљиве структуре, тј. документи. Током рада, откривају се удели различитих тема по документима као и расподеле речи унутар тема. Важно је напоменути да **именовање** тема не постоји у основној верзији алгоритма. Алгоритам групише речи у одређене целине - теме, а наслovљавање тема се препушта стручњацима.

Нека је дат једноставан документ који, након склањања везника, личних заменица и осталих шумова (енг.*stopwords*) садржи речи приказане у следећој табели.

Etruscan	trade	price	temple	market

Слика 2.3: Преузето са [?]

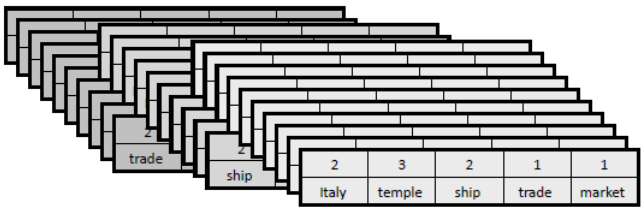
Процес моделовања тема започоње **случајним додељивањем** тема свакој од речи у документу. Дакле, пошто не постоји никакво знање о присуству тема у документима као ни о томе која реч припада којој теми, ову доделу је неопходно урадити на случајан начин. Интуитивно је јасно да се за тако нешто унапред мора одредити број тема који се захтева у задатом скупу докумената. Више о улазним параметрима алгоритма може се наћу у одељку 3 овог рада. Пример једне случајне доделе дат је на следећој слици .

3	2	1	3	1
Etruscan	trade	price	temple	market

Слика 2.4: Преузето са [?]

На овај начин је направљена иницијална **расподела** тема унутар посматраног документа - 40% текста говори о теми 3, 40% о теми 1, док 20% говори о теми 2.

Уколико сада на сличан начин доделимо теме и осталим документима, полазни скуп докумената може се приказати следећом сликом.



Слика 2.5: Преузето са [?]

Пошто сваки документ има иницијалну, **случајну** расподелу тема, једноставно је груписати речи унутар тема и на тај начин направити иницијалну **случајну** расподелу речи по темама. Одређивање расподеле по темама, може се илустровати следећом сликом.

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Слика 2.6: Преузето са [?]

У првој колони табеле уписане су све речи из свих докумената (речник) док је затим у свакој од наредних колона уписан одговарајући број који означава колико пута је дата реч додељена тој теми. Рецимо, бројеви 1 0 35 у другој врсти табеле означавају да је реч *Etruscan* 35 пута била сврстана у тему број 3, једном је била додељена теми број 1 док се ниједним није нашла у теми број 2.. Дакле, почевши од друге колоне приказане табеле, табела по колонама, садржи *расподелу речи по темама*. Једноставним сортирањем колона, добијају се највероватније речи у свакој од тема.

У овом тренутку, расподеле које су добијене нису релевантне зато што у позадини стоји апсолутно случајно додељивање тема које није базирано на документима, тј. на јединим видљивим подацима. Дакле, потребно је добијене резулатате *прилагодити* тако да осликавају тематску структуру документа. Прилагођавање се одвија у одређеном, унапред познатом броју итерација. Генерално, што је већи број итерација, то су добијене расподеле релевантније, мада, како резултати показују, постоје и нека ограничења за ове вредности. Више о одређивању оптималног броја итерације може се наћи у одељцима 3 и 6 овог рада.

Унутар једне итерације, за сваки документ за сваку реч унутар тог документа врши се провера колико је тренутно додељена тема адекватна, тј. да ли постоји боља тема којој би та реч могла бити додељена. На тај начин, из итерације у итерацију, расподеле све више и више осликавају структуру докумената.

Нека је почетна расподела по темама дата на претходним сликама (2.5,2.6). Нека се провера подобности теме прво извршава за реч *trade* првог документа. Ова реч је унутар првог документа додељена теми број 2, док је, гледано са становишта свих докумената који се посматтају, укупно 8 пута сврстана у ту тему. Потребно је испитати да ли тема број 2 највише одговара тој речи. Претпоставимо да знамо теме за све остале речи, како из документа који се посматра, тако и за остале документе, и да је једино непозанто којој теми припада *trade* у посматраном документу. Дакле, расподела тема у посматраном документу као и расподела речи по темама, сада изгледа као на следећој слици и потребно је доделити изабраној речи тему унутар посматраног документа.

3	?	1	3	1
Etruscan	trade	price	temple	market

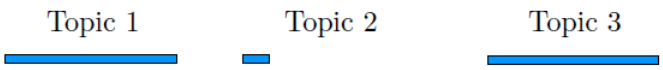
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	**7**	1
...			

Слика 2.7: Преузето са [?]

Уколико се посматра расподела тема у изабраном документу, приметиће се да документ највише "говори о"темама 3 и 1 док о теми 2 не говори уопште. Према томе, удели тема 3 и 1 су значајни, док је удео теме 2 занемарљиво мали. Обзиром да је основна претпоставка овог модела да сви документи говоре о свим темама, не може се рећи да изабрани документ уопше "не говори"о теми 2. Начин на који ће се означити да је тема 2 јако слабо присутна у изабраном документу врши се тако што се теми 2 додели јако мали удео. Обизиром да је у питању реч из документа који има неки одређену расподелиу тема, логично је очекивати да избор теме за ту реч зависи од тема тог документа.

Удели тема у изабраном документу могу се представити дужином линија, као што је приказано на следећој слици.

3	?	1	3	1
Etruscan	trade	price	temple	market



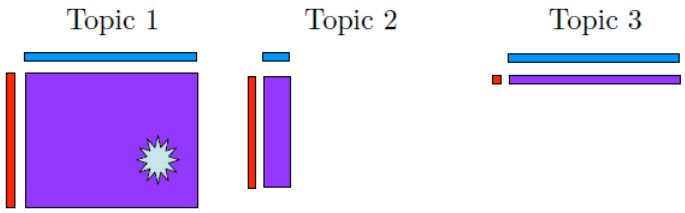
Слика 2.8: Преузето са [?]

Међутим, како је речник (скуп речи) заједнички за све документе, тема речи зависи и од глобалног присуства те речи у свим темама. И ова претпоставка је логична, јер, као што је на почетку наведено, свака реч припада свим темама, само са различитим верпватноћама. Глобално присуство изабране речи у свим темама приказано је на следећој слици :

	1	2	3
trade	10	7	1

Слика 2.9: Преузето са [?]

Дакле, избор теме за реч *trade* зависи од расподеле тема у посматраном документу као и од присуства те речи у свим темама. Ова зависност може се представити следећом сликом :



Слика 2.10: Преузето са [?]

Вертикална, црвена линија представља присуство речи у одговарајучој теми (формалније, вероватноћу са којом се та реч налази у изабраној теми). Љубичаста "површина"представља подобност да одговарајућа тема буде додељена тој речи у изабраном документу. Како се слике јасно може уочити, "највећу"површину формира тема 1 те је, према томе, речи *trade* додељује тема број 1 у овој итерацији. Након извршене измене, расподела тема у документу, као и расподела речи по темама, приказана је на следећој слици :

3	**1**	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	**11**	7	1
...			

Слика 2.11: Преузето са [?]

На овај начин, избараној речи додељена је најверпватнија тема а тематска слика документа "више личи"на реалну слику.

Ако се описани процес примени на сваку реч сваког документа, расподеле ће из итерације у итерацију све више осликавати структуру полазних докумената. Обзиром да избор теме за сваку реч зависи од тренутно додљених расподела тема и речи, овим процесом се узимају у обзир видљиви подаци. Дакле, скривене структуре (непознате расподеле) се генеришу на основу једино видљивих података - докумената и речи у њима.

Узевши све претходно у обзир, јасно је због чека не постоји именовање тема. Када се процес "генерисања"заврши (нпр. достигне одређени број итерација), сортирањем одговарајућих колона из табеле са слике 2.7, добија се расподела речи по темама. На основу експретског знања овим расподелама додељују се имена - нпр. тема 1 - математика, тема 2- економија итд.

Глава 3

Математичка позадина

У претходним поглављима, рад се углавном бавио питањима *шта је TM алгоритам и чему служи*, без улажења у то ***како*** он уствари ради.

Опис рада TM алгоритама - конкретно LDA имплементације, биће изложен у неколико целина. Најпре ће се објаснити (увести) неки појмови вероватноће који су битни за разумевање сусштине рада алгоритма, а затим ће бити изнешена математичка позадина самог алгоритма.

3.1 Теорија вероватноће

Теорија вероватноће је математичка дисциплина која се бави изучавањем случајних појава тј. појава чији исходи нису увек строго дефинисани.

Први проблеми који се могу сматрати проблемима вероватноће потичу још из 12. века и везани су за проучавање исхода разних игара на срећу. Развој *теорије вероватноће* почиње средином 17. века и везан је за имена Блеза Паскала,Пјера де Ферма и Кристијана Хајгенса. Наиме, између Паскала и Ферма је 1654. године започела интересантна преписка о низи проблема међу којима је био и проблем везан за поделу улога приликом прекида једне коцкарске игре. Проблем је био постављен на следећи начин : Два играча А и Б се договоре да читав улог припадне ономе ко први добије три игре. Када је играч А добио 2 игре а играч Б једну игру, играчи су споразумно одлучили да прекину игру. Поставља се питање како сада да поделе улог. Паскал је предложио поделу у односу 3:1 у корист играча А. Овај пример често се узима као почетак настанка теорије вероватноће.

Неке од појава које се догађају у реалном свету лако се могу предвидети и објаснити услод познавања законитости њиховог настанка. У такве појаве спадају нпр. помарачење Сунца и Месеца, плима и осека, гравитација итд. Међутим, постоје појаве чије узроке тренутно није могуће одредити па се не могу у потпуности објаснити и одредити. Неке од таквих појава су нпр. добитак на лутрији или метереолошке појаве. Прилоком бацања металног, хомогеног новчића, никада није сигрно да ли ће пасти писмо или глава. Међутим, уколико бацамамо новчић много пута, може се уочити да је отприлике исти број пута пало писмо као и глава (такве експерименте су радили Буфон и ПирсонДакле, законитост код оваквих догађаја може се уочити тек након великог броја понављања појаве.

3.1.1 Основни појмови

Основни полазни појам у теорији вероватноће је непразан скуп Ω који представља скуп свих могућих исхода једног експеримента. Овај скуп се често назива **простор елементарних догађаја** и може бити коначан, пребројив или непребројив. **Случајни догађаја** или само **догађај** представља било који подскуп скупа Ω. Најчешће се случајни догађаји означавају великим, штампаним, латиничним словима. За догађај **A** каже се да се **реализовао** ако

се реализовао неки исход ω који припада скупу А. Догађај који је садржи све могуће елементарне исходе експеримента назива се **сигуран догађај** а догађај који не садржи ни један елементарни исход назова се **немогућ догађај**.

Пример: Нека је дата хомогена коцка чије су стране означене бројевима од 1 до 6. Елементарни догађаји су појављивање одређеног броја при бацању коцкице. Према томе, скуп свих могућих исхода екперимента бацања коцкице је $\Omega = \{1, 2, 3, 4, 5, 6\}$. Догађај А = "пао је паран број"одређује скуп А= $\{2, 4, 6\}$

Производ два догађаја и , у ознаци AB је догађај који се реализује ако и само се ако реализују оба догађаја. Дакле, производ догађаја је пресек скупова А и В. Уколико су А и В дисјунктни скупови (пресек је празан скуп) за такве догађаје кажемо да су **несагласни** или да се **искључују**.

Збир два додгађаја А и В, у ознаци $A \cup B$ представља догађај који се реализује ако се реализује бар један од догађаја А и В.

Разликом догађаја А и В, у ознаци $A - B$ назива се догађај који се реализује ако и само ако се реализује догађај А а не реализује догађај В.

Потпун систем догађаја : За догађаје $A_1, A_2, ..A_n$ се каже да образују *потпун систем догађаја* уколико важи : $\bigcup_i A_i = \Omega$. Дакле, при реализацији неког експеримента бар један од догађаја $A_1, A_2, ..A_n$ ће се реализовати. Посебно су интересантни потпуни системи несагласних догађаја као што се може видети код формуле тоталне вероватноће.

Дефиниција 1 (Класична дефиниција вероватноће) *: Нека је $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$ скуп свих могућих једнаковероватних елементарних догађа који су међусобно несагласни и нека је $A = \{\omega_{i_1}, \omega_{i_2}, ..., \omega_{i_m}\}$ догађај који се састоји од m елементарних једнаковероватних догађаја. Вероватноћа наступања догађаја A је :*

$$P(A) = \frac{m}{n} \tag{3.1}$$

Претходна дефиниција може се неформално изразити и овако : вероватноћа догађаја А једнака је количнику броја **повољних исхода** експеримена (исходи када се реализује догађај А) и укупног броја свих могућих исхода експеримента.

Класична дефиниција вероватноће је применљива само онде где су елементарни догађаји једнаковероватни. Међутим, тај услов је у пракси јако тешко испунити. Чак и у случајевима када је то наизглед очигледно, као што је бацање коцкице, једнаковероватност не може бити гарантована. Разлози за то могу бити технологија израде коцкице која не мора бити савршено прецизна, немогућност обезбеђивања идеалних и непромењљивих услова током извођења експеримента итд. Због тога је једини начин којим је могуће заиста утврдити вероватноћу догађаја А *статистички приступ* заснован на великом броју експеримената.

Дефиниција 2 (Статистичка дефиниција вероватноће) *: Нека се у n понављања експеримента изведених под приближно истим условима догађај A реализовао m_n пута. Вероватноћа догађаја A је*

$$P(A) = \lim_{n \to \infty} \frac{m_n}{n} \tag{3.2}$$

3.1.2 Условна вероватноћа

Вероватноћа догађаја чија реализација **не зависи** од наступања било ког другог догађаја назива се **безусловна вероватноћа**. Ако је реализација догађаја А условљена реализацијом неког догађаја В при чему В није немогућ догађај ($P(B) \neq 0$), тада се вероватноћа догађаја под условом да се десио догађај В назива **условном вероватноћом** и означава се са $P(A \mid B)$. Дакле, $P(A \mid B)$ је вероватноћа догађаја А под условима који сигурно доводе до реализације догађаја В.

Нека се изводи експеримент у коме постоји n једнаковероватних елементарних догађаја и нека је са n_A, n_B, n_{AB} означен број елеменгтарних догађаја који доводе до реализације догађаја A, B, AB редом.

Према класничној дефиницији вероватноће, вероватноћа реализације догађаја A и AB је :

$$P(B) = \frac{n_B}{n}, P(AB) = \frac{n_{AB}}{n} \tag{3.3}$$

Ако је реализација догађаја A условљена реализацијом догађаја B, то је број повољних исхода догађаја A n_{AB} (број елементарних догађаја који имају осбине и скупа A и скупа B). Пошто се догађај A реализује само ако се реализовао догађај B, број свих могућих исхода је n_B (број свих могућих елементарних догађаја када наступа догађај B). Дакле, условна вероватноћа догађаја A, под условом да се десио догађај B је :

$$P(A \mid B) = \frac{n_{AB}}{n_B} = \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n}} = \frac{P(AB)}{P(B)} \quad , P(B) \neq 0 \tag{3.4}$$

У случају да је догађај B условљен догађајом A, аналогно се изводи да је

$$P(B \mid A) = \frac{P(AB)}{P(A)} \quad , P(A) \neq 0 \tag{3.5}$$

Из релација (3.4) и (3.5) следи

$$P(AB) = P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \tag{3.6}$$

Релација (3.6) назива се још и **теорема о производу вероватноћа**

Теорема 1 (Формула тоталне вероватноће) *: Ако су $H_1, H_2, .., H_n$ међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, .., n)$ при чему важи $H_1 + H_2 + .. + H_n = \Omega$ тада је :*

$$P(A) = \sum_{i=1}^{n} P(H_i) P(A \mid H_i) \text{ за сваки догађај} A \subseteq \Omega \tag{3.7}$$

Напомена : Догађаји $H_1, H_2, .., H_n$ чине потпун систем несагласних догађаја.

Доказ 1 *Обзором да су догађаји подскупови скупа свих елементарних догађаја очигледно је да важи*

$$A = A\Omega = A \sum_{i=1}^{n} H_i = \sum_{i=1}^{n} AH_i. \tag{3.8}$$

На основу релације (3.6) следи :

$$P(A) = P(\sum_{i=1}^{n} AH_i) = \sum_{i=1}^{n} P(AH_i) = \sum_{i=1}^{n} P(H_i) P(A \mid H_i) \tag{3.9}$$

Вероватноће $P(H_i)$ су обично познате унапред и називају се **априорним вероватноћама** а сами догађаји **хипотезама**.

Теорема 2 (Бајесова формула [1]) *: Ако су $H_1, H_2, .., H_n$ међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, .., n)$ при чему важи $H_1 + H_2 + .. + H_n = \Omega$ тада је :*

$$P(H_i \mid A) = \frac{P(H_i) P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j) P(A \mid H_j)} \quad (i = 1...n) \quad \text{за сваки догађај} A \subseteq \Omega \tag{3.10}$$

Доказ 2 *Из релације (3.6) следи :*

$$P(H_iA) = P(H_i)P(A \mid H_i) = P(A)P(H_i \mid A) \quad (i = 1...n) \tag{3.11}$$

Условна вероватноћа догађаја H_i под условом да се десио догађај A је:

$$P(H_i \mid A) = \frac{P(H_iA)}{P(A)} = \frac{P(H_i)P(A \mid H_i)}{P(A)}$$

Примењујући формулу потпуне вероватноће за $P(A)$ добија се

$$P(H_i \mid A) = \frac{P(H_i)P(A \mid H_i)}{\sum_{j=1}^{n} P(H_j)P(A \mid H_j)}$$

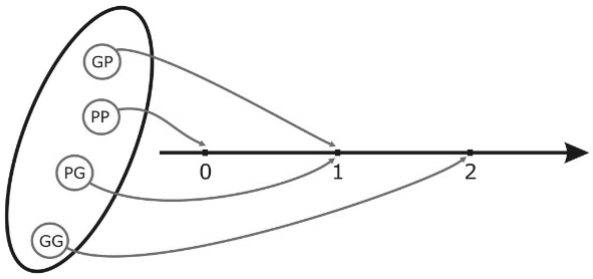
што представља Бајесову формулу.

3.1.3 Случајне променљиве

Ако се сваком елементарном догађају придружи један реалан број, онда се извођење експеримента може посматрати као избор вредности једне променљиве. Променљива величина која те бројене вредности узима са одређеним вероватноћама назива се *случајна променљива*. Дакле, уместо вербалне карактеризације догађаја (описа речима шта догађај представља) много је једноставније за рад догађаје окарактерисати бројним вредностима тј. неким реалним бројевима.

Пример 1 : У експериманту бацања новчића могућа су два елементарна исхода : грб или писмо. Нека је догађај који се посматра "пало је писмо". Појава писма се може означити бројем 1 а појава грба бројем 0. Сада се овај експеримент може замислити као избор 0 или 1 са вероватноћом $\frac{1}{2}$

Пример 2 : Новчић се баца два пута. Нека је са P означена појава писма а са G појава грба. Скуп свих елементарних исхода експеримента је $\Omega = \{PP, PG, GP, PP\}$. Нека је догађај који се посматра "број палих писама". Сваком исходу се може доделити један реалан број и то $PP- > 2, GP- > 1, PG- > 1, GG- > 0$. Ово додељивање вредности се карактерише случајном променљивом. Случајна променљива сваку од ових вредности узима са различитом вероватнићом.



Слика 3.1: Графички пример случајне променљиве

Дефиниција 3 *Функција X која сваком случајном догађају $\omega \in \Omega$ додељује неки реалан број $X(\omega)$ назива се* ***случајна променљива*** *где је $X : \Omega \longrightarrow R$*

Дакле, случајна променљива је пресликавање скупа Ω у скуп **реалних** бројева за разлику од вероватноће која је пресликавање скупа Ω у скуп $[0, 1]$

Важно је уочити да случајна променљива **нема конкретну вредност** већ се само говори о вероватноћама да узме неки конкретну вредност.

Разликују се два основна типа случајних променљивих - **дискретне** и **непрекидне**. Подела се врши у зависности од тога да ли случајна променљива узима вредности у пребројивом или непребројивом скупу вредности.

3.1.3.1 Дискретне случајне промењиве

За случајну промењиву се каже да је дискретног типа ако узима коначан број изолованих вредности или пребројиво много вредности

Дефиниција 4 *Нека случајна промењива X може да узме вредности $x_1, x_2, ..., x_n$ са вероватноћама $p_1, p_2, ..., p_n$ при чему важи да је $p_1 + p_2 + ... + p_n = 1$. Скуп парова $(x_i, p_i = P\{X = x_i\}), i = 1, 2, ..., n$ или написано :*

$$\begin{pmatrix} x_1 & x_2 & \cdots \\ p(x_1) & p(x_2) & \cdots \end{pmatrix}$$

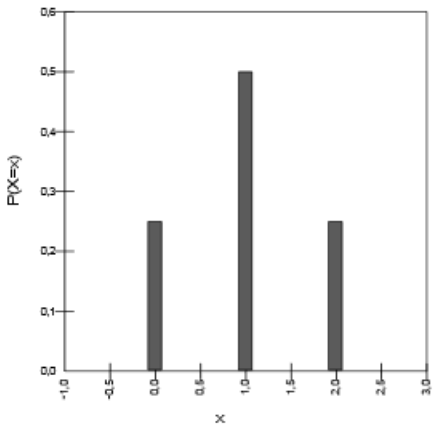
чине закон расподеле или распоред вероватноћа случајне промењиве X.

Закон расподеле случајне промњиве може да се посматра као **правило** по коме се свакој вредности случајне промењиве придружује одговарајућа вероватноћа. Дакле, при реализацији експеримента сигурно ће се десити догађај којем је придужена нека вредност случајне промењиве. Због тога је сума свих вероватноћа у расподели случајне промењиве 1. Међутим, нису све вредности подједнако вероватне па се свакој вредноти придружује вероватноћа са којом се очекује. Претходна дефиниција може се интерпретирати и на следећи начин : извесна маса једнака јединици је распоређена на такав начин да се у тачкама $x_1, x_2, ..., x_n$ налазе одговарајући делови масе $p_1, p_2, ..., p_n$. Услед оваквог тумачења, закон расподеле вероватноћа се често назива и **функција масе вероватноћа**

У примеру 2, случајна промењива може да узме три вредности, тј. писмо се може појавити 0,1 или 2 пута у два бацања. Ни један други исход није могућ - нпр. у два бацања писмо не може да се појави 3 пута или -1 пут. Међутим, вероватноћа да се писмо неће појавити ни једном (или да се појави два пута) је $\frac{1}{4}$ - вероватноћа да падне глава у првом бацању је $\frac{1}{2}$ и вероватноћа да падне глава у другом бацању је $\frac{1}{2}$, дакле, вероватноћа да оба пута падне глава је је $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, вероватноћа да се писмо појави тачно једном је $\frac{1}{2}$ - писмо пада тачно једном у случају PG или GP. Вероватноћа за оба ова догађаја је $\frac{1}{4}$. Дакле, вероватноћа да се десио бар један од ових догађаја је $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Према томе, расподела случајне промењиве "број појављивања писма у два бацања "је :

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Закон расподеле дискретне случајне промењиве може се представити графички, као на следећој слици :



Слика 3.2: Графички пример случајне промењиве

На апсциси се налазе могуће вредности случајних променљивих док се на ординати налазе вероватноће са којом случајна променљива узима дату вредности. Са претходне слике јасно се уочава дискретност посматране случајне променљиве - вероватноћа да случајна промељива узме вредност између неке две целобројне вредности је 0.

Функција расподеле дискретне случајне променљиве:

Распоред или закон расподеле случајне променљиве дискретног типа може се представити као листа свих могућих вредности случајне променљиве и одговарајућих вероватноћа. Међутим, поставља се питање како представти случајну променљиву која може узимати јако пуно вредности тј. бесконачно много вредности. У овом случају би требало формирати листу од бесконачно много чланова, што је практино неизводљиво. (Пример једне такве случајне променљиве би био - број бацања коцкице док се не добију две узастопне шестице. Случајна променљива може узети вредности 2,3,4,... са различитим вероватноћама, при чему не постоји горња граница броја бацања при којој се сигурно добијају две узастопне шестице). Због описаног проблема, потребно је пронаћи другачији начин представљања случајне променљиве и одговарајућих вероватноћа. То се постиже **функцијом расподеле** која се може дефинисати за сваку случајну променљиву.

Дефиниција 5 ***Функција расподеле*** *(још се назива и* ***кумулативна функција расподеле****) дискретне случајне променљиве претставља вероватноћу да случајна променљива X узме вредност која је мања или једнака неком реалном броју x при чему је дефинисана за свако реално x.*

$$F(x) = P(X < x) \quad \forall x \in R$$

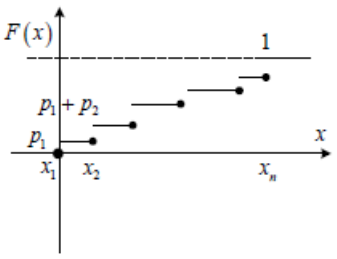
Дакле, кумулативна функција расподеле има облик

$$F(x) = \begin{cases} 0, & x \leq x_1 \\ p_1, & x_1 < x \leq x_2 \\ p_1 + p_2, & x_2 < x \leq x_3 \\ .... & ... \\ 1 & x > x_n \end{cases}$$

и може се изразити као :

$$F(x) = \sum_{k, x_k < x} P(X = x_k)$$

а графички приказ је дат на следећој слици :



Слика 3.3: График кумулативне функције расподеле
случајне променљиве дискретног типа

Две најважније дискретне расподеле су **Биномна** и **Пуасонова** расподела.

3.1.3.2 Непрекидне случајне промењиве

Случајна промењива је (апсолутно) непрекидног типа ако може да узме **било коју** вредност из неког интервала. Број вредности које може да узме случајна промењива непрекидног типа је **бесконачан**. Неки од примера су : висина и тежина људи, дужина трајања батерије итд. На пример, нека је X случајна промењива која представља дужину рада сијалице. Ова случајна промењива може узети било коју вредност на интервалу од 1 до нпр. 1000 сати. Како у интервалу $[0, 1000]$ има бесконачно много реалних бројева, не постоји начин да се дефинише вероватноћа за сваку појединачну вредност, као што је био случај код дискретних промењивих. Такође, интуитивно је јасно да је вероватноћа да ће сијалица прегорети у тачно одређеном тренутку једнака 0 док је вероватноћа да ће прегорети у неком временском интервалу различита од нуле.

Дефиниција 6 *Случајна промењива X је апсолутно непрекидног типа ако постоји* ***ненегативна*** *функција* $f : \mathbb{R} - > \mathbb{R}$ *таква да за било који интервал* $[a, b] \subset (-\infty, \infty)$ *важи :*

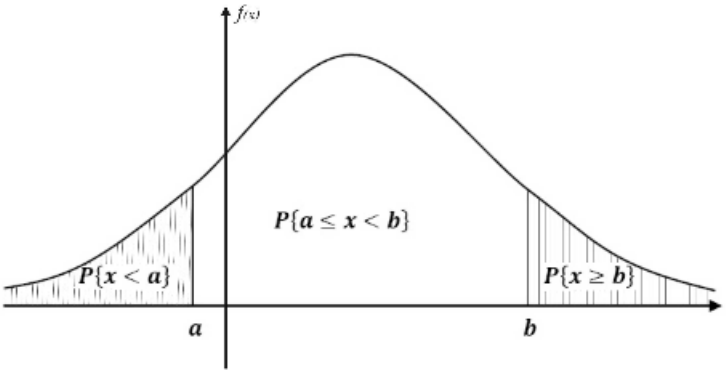
$$P\{a \leq X < b\} = \int_a^b f(x)dx \tag{3.12}$$

Функција $f(x)$ мора да задовољи услов :

$$P\{-\infty \leq X < \infty\} = P\{\Omega\} = \int_{-\infty}^{\infty} f(x)dx = 1$$

Функција $f(x)$ се назива **густина расподеле вероватноће** случајне промењиве X. Случјане промењиве **дискретног типа** немају густину расподеле баш као што ни случајне промењиве непрекидног типа немају закон расподеле вероватноћа.

Из релације (3.12) следи да је вероватноћа да случајна промењива узме вредност из интервала $[a, b]$ једнака **површини** испод графика функције $f(x)$ на интервалу $[a, b]$.



Слика 3.4: Функција густине

Функција расподеле непрекидне случајне промењиве:

Дефиниција 7 ***Функција расподеле (кумулативна функција расподеле)*** *непрекидне случајне промењиве се може представити као :*

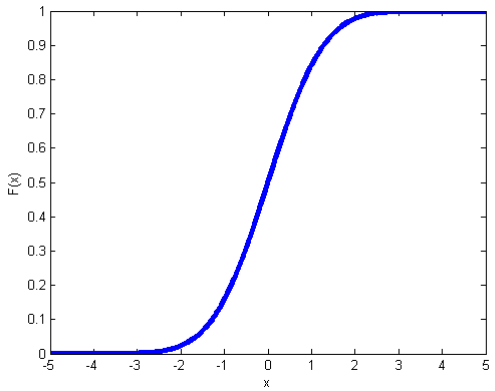
$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt) \quad x \in (-\infty, \infty)$$

где је $f(x)$ *функција густине.*

Дефиниција кумулативне суме преко интеграла је јаснија ако се има на уму интервал из ког случајна променљива може да узме вредности. Код случајних променљивих дискретног типа, тај скуп је био пребројив па се кумулативна функција расподеле дефинисала преко суме. Случајне променљиве непрекидног типа могу узти бесконачно много вредности па се сума код дискретних случајних променљивих (када број тачака тежи у бесконачност) замењује интегралом.

Напомена : Ако случајна променљива X не узима све вредности из интервала $(-\infty, \infty)$ усваја се да је $f(x) = 0$ за све вредносто x из интервала у којима X не узима вредности.

График кумулативне функције расподеле непрекидне случајне променљиве X је сада представљен глатком кривом линијом (за разлику од случајне променљиве дискретног типа где је график био "степенаст").



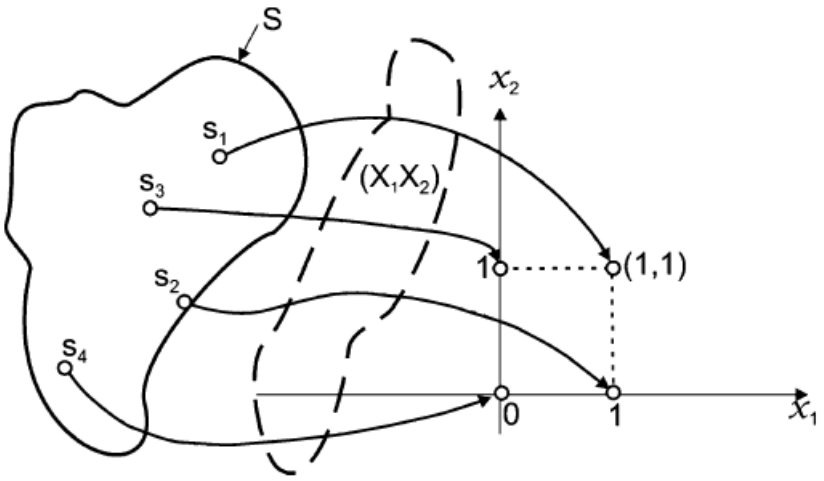
Слика 3.5: Кумулативна функција расподеле
за случајне променљиве непрекидног типа

3.1.3.3 Вишедимензионалне случајне прменљиве

Случајна променљива представља пресликавање скупа догађаја у реалне бројеве. Дакле, излази експеримента се мапирају у једнодимензионалан простор реалних бројева. Међутим, постоје случајеви када је потребно излазе експеримента мапирати у вишедимензионалне реалне просторе. На пример, при истовременом бацању два новчића могућа су 4 исхода :

1. s_1: први новчић писмо - други новчић писмо
2. s_2: први новчић писмо - други новчић глава
3. s_3: први новчић глава - други новчић писмо
4. s_4: први новчић глава - други новчић глава

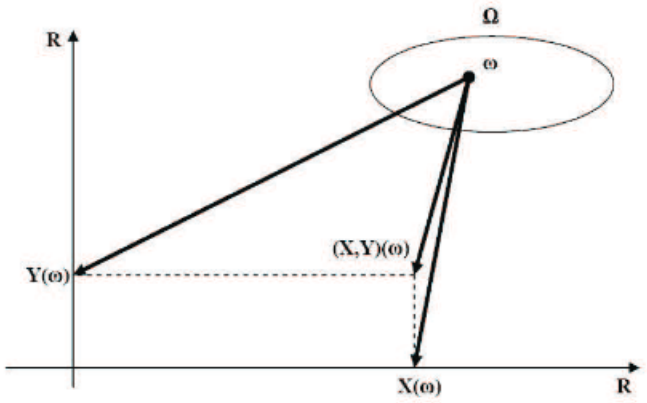
Нека је са X_1 означена случајна променљива која узима вредност 1 ако се на првом новчићу појавила глава, односно 0 ако се појавило писмо и аналогно X_2 која на исти начин означава појаву главе на другом новчићу. Исход експеримента се сада може описати дводимензионалном променљивом (X_1, X_2). Графички приказ ове дводимензионалне промељиве дат је на следећој слици :



Слика 3.6: Експеримент : бацање два новчића.
Скуп S представља скуп свих елементарних исхода (Ω)

Дефиниција 8 *Ако су* $X : \Omega -> \mathbb{R}, Y : \Omega -> \mathbb{R}$ *случајне променљиве, тада се* ***уређени пар*** (X, Y) *натива* ***дводимензионална случајна променљива****. Уређеним паром* (X, Y) *се сваком исходу* $\omega \in \Omega$ *придружује уређени пар бројева* $(X(\omega), Y(\omega)) = (x, y) \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

На следећој слици графички је представљена дводимензионална случајна променљива.



Слика 3.7: Дводимензионална случајна променљива

Овако уведен појам дводимензионалне случајне променљиве се може проширити и на више димензија и тада настају n-димензионалне случајне променљиве. Закључци изведени за дводимензионалне се такође односе и на вишедимензионалне случајне приомењиве.

Кумулативна функција расподеле (још се назива и *заједничка расподела енг. joint distribution* дводимензионалне случајне променљиве, у ознаци $F_{XY} : \mathbb{R}^2 -> [0, 1]$ дефинише се као вероватноћа реализације догађаја $\{X \leq x, Y \leq y\}$ односно :

$$F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\} \quad -\infty < x, y < \infty$$

Неке карактеристике функције расподеле дводимензионалне случајне променљиве :

1. $0 \leq F_{X_1,X_2}(x_1, x_2) \leq 1$
2. $F_{X_1,X_2}(-\infty, -\infty) = 0$

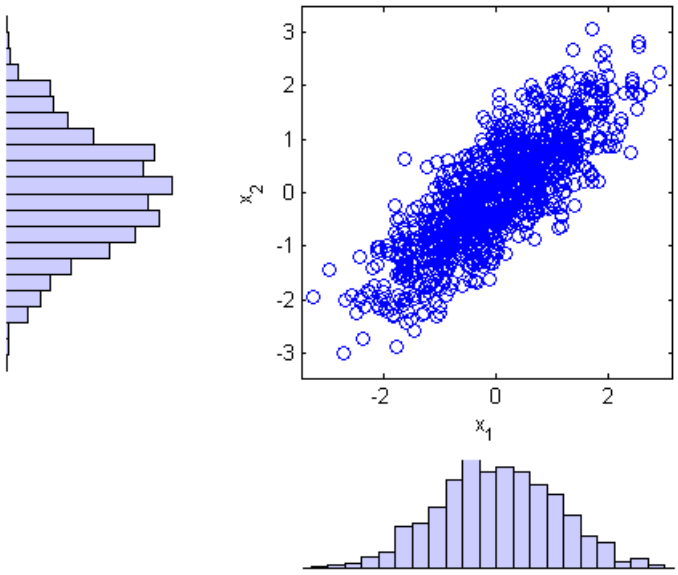
3. $F_{X_1,X_2}(-\infty,-\infty)=0$
   $F_{X_1,X_2}(-\infty,x_2)=0$ $F_{X_1,X_2}(x_1,-\infty)=0$

4. $F_{X_1,X_2}(\infty,\infty)=1$

5.
$$F_{X_1,X_2}(x_1,\infty)=F_{X_1}(x_1) \tag{3.13}$$

6.
$$F_{X_1,X_2}(\infty,x_2)=F_{X_2}(x_2) \tag{3.14}$$

Једнакостима (3.13) и (3.14) су дефинисане **маргиналне расподеле** случајних променљивих X_1 и X_2. Маргиналне расподеле су уствари расподеле једнодиментионалних случајних променљивих X_1 и X_2. На следећој слици је предсављена заједничка расподела две случајне приомењиве заједно са њиховим маргиналним расподелама.

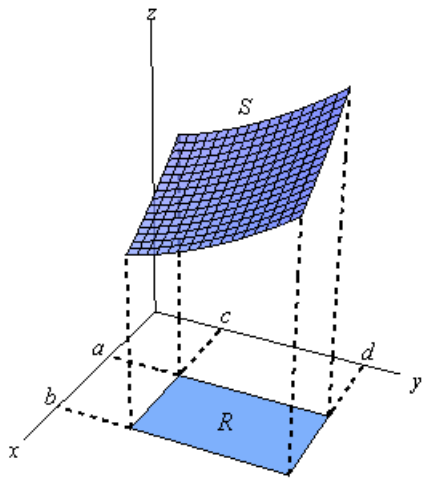


Слика 3.8: Дводимензионална случајна променљива и маргиналне расподеле

Заједничка функција густине (енг. joint density function) дводимензионалне случајне променљиве се дефинише као :

$$f_{X_1,X_2}(x_1,x_2)=\frac{d^2F_{X_1,X_2}(x_1,x_2)}{dx_1dx_2} \tag{3.15}$$

У случају једнодимензионалне случајне променљиве, површина испод графика функције густине на неком интервалу представљала је вероватноћу да случајна променљива узме вредност из тог интервала. У случају дводимензионалне случајне промељиве од интереса је пронаћи вероватноћу да она узме вредност из неке **области**. Та вероватноћа представља **запремину** тела ограниченог функцијом густине са горње стране и датом облашћу са доње стране.



Слика 3.9: Вероватноћа да дводемнзионална случајна промељива (X, Y) узме вредности из области R

Неке особине функције густине :

1. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2 = 1$

2. $F_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

3. $F_{X_1}(x_1) = \int_{-\infty}^{x_1}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$
   $F_{X_2}(x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

4. $f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_2$
   $f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1,x_2)dx_1$

5. $P\{x_{11} < X_1 \leq x_{12}, x_{21} < X_2 \leq x_{22}\} = \int_{x_{21}}^{x_{22}}\int_{x_{11}}^{x_{12}} f_{X_1,X_2}(x_1,x_2)dx_1dx_2$

Функција условне расподеле и густине

У неким специфичним случајевима је потребно пронаћи расподелу једне случајне променљиве знајући вредност друге случајне променљиве. Таква расподела назива се **условном расподелом** и обележава се са $F_{X_1}(x_1 \mid X_2 = x_2)$. Аналогно, може се дефинисати и проблем налажења функције густине једне случајне променљиве знајући вредност друге случајне променљиве и таква функција густине се означава са $f_{X_1}(x_1 \mid X_2 = x_2)$. Према [**?**] условна расподела односно густина се рачуна по следећем обрасцу (детаљно извођење се може наћи у [**?**])

$$F_{X_1}(x_1 \mid X_2 = x_2) = \frac{\int_{-\infty}^{x_1} f_{X_1,X_2}(x_1,x_2)dx_1}{f_{X_2}(x_2)}$$

односно :

$$f_{X_1}(x_1 \mid X_2 = x_2) = \frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)}$$

3.2 Важније расподеле

3.2.1 Бинонма и полиномна(енг. multinomial) расподела

Бинонма расподела

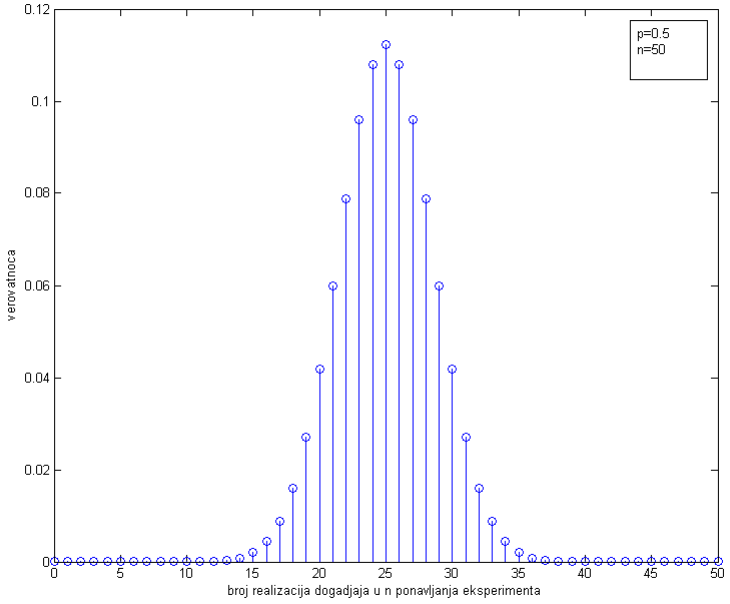
Нека ја A догађај неког екеперимента E који се реализује са вероватноћом $P(A) = p$. Тада је вероватноћа супротног догађаја $P() = 1 - p = q$. Резултат експеримента који је до интереса је остваривање или неостваривање догађаја A. Дакле, може се сматрати да је скуп свих елементарних исхода $\Omega = \{A, \overline{A}\}$. Нека се експеримент понавља **независно** и у неизмењеним условима n пута. На тај начин је формиран **сложени експеримент** чији скуп елементарних исхода садржи све могуће n-торке састављене од и $\overline{A}$ и има их укупно 2^n. Нека је, даље, на том скупу елементарних исхода дефинисана случајна променљива X_n као број остваривања догађаја A у n понављања експеримента E. Вероватоћа да ова случајна променљива узме конкретну вредност k је :

$$p_k = P\{X_n = k\} = \binom{n}{k} p^k q^{n-k}$$

Вероватноће $P\{X_n = k\}, (k = 0, 1, .., n)$ дефинишу **биномну расподелу**, у ознаци $\mathbb{B}(n, p)$. Ова расподела је дискретног типа а њена функција расподеле(кумулативна) се може изразити као :

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ \sum_{k=0}^{r} \binom{n}{k} p^k q^{n-k} & 0 < r < n \\ 1 & x > n \end{cases}$$

Закон расподеле вероватноћа случајне променљиве биномне расподеле приказан је на следећој слици :



Слика 3.10: Биномна расподела - закон расподеле

Полиномна (енг. multivariate) расподела

Изводи се серија од n независних експеримената при чему резултат експеримента може бити један од **коначно много** догађаја : $A_1, A_2, .., A_k, \sum_{i=1}^{k} A_i = \Omega, P(A_i) = p_i (i = 1, 2, ..k)$. Ако се дефинише k-диментионална случајна применљива $(S_n^{(1)}, ..., S_n^{(k)})$, где $S_n^{(i)}$ представља број релаизација случајног догађаја A_i у n независних експеримената, тада важи :

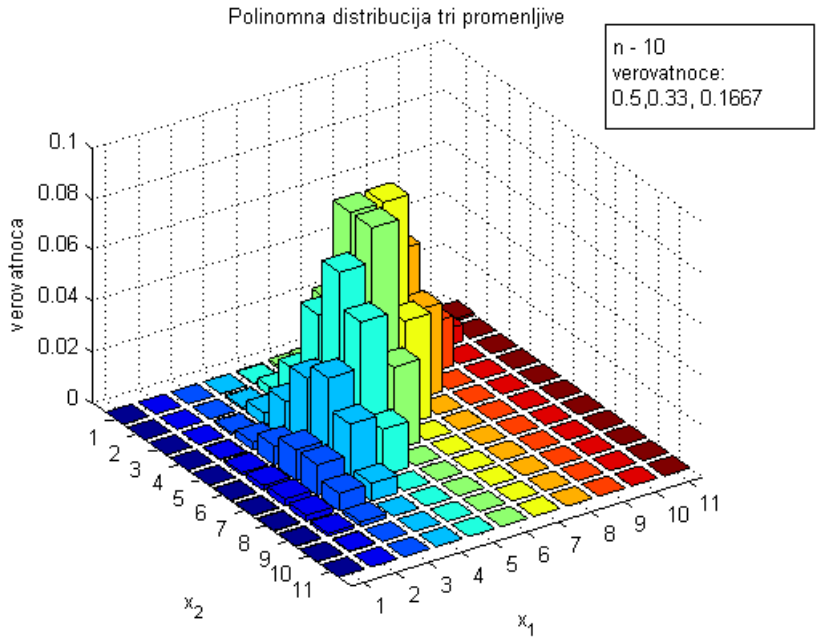
$$P(S_n^{(1)} = r_1, .., S_n^{(k)} = r_k) = \frac{n!}{r_1!..r_k!} p_1^{r_1}...p_k^{r_k}$$

$$r_1, ..r_k \in \{0, 1, .., n\} \quad r_1 + .. + r_k = n$$

Ако се са $S = (S_n^{(1)}, ..., S_n^{(k)})$ означи k-диментионална случајна применљива која има полиномијалну расподелу тада се то записује као :

$$S \sim Mult(n, p)$$

где је $p = (p_1, p_2, .., p_k)$ Пример полиномне расподеле при чему резулатат експеримента може бити један од **три** догађаја, дат је на следећој слици :



Слика 3.11: Биномна расподела - закон расподеле

3.2.2 Дирихлеова расподела

Дирихлеова расподела представља фамилију расподела **за параметре** p полиномијалне расподеле. Задаје се са :

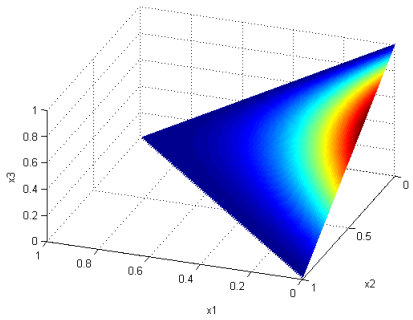
$$Dir(p; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}$$

при чему је α параметар расподеле а B означава мултиномијалну *бета функцију*. Мултиномијалну бета функција се изражава преко гама фунцкије на следећи начин

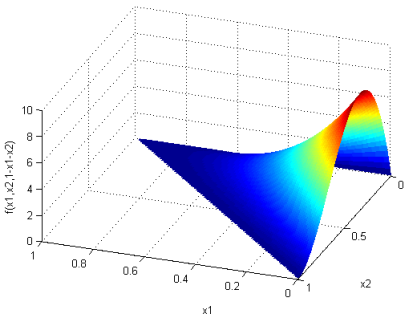
$$B(\alpha) = \frac{\prod_{i=1}^{|\alpha|} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{|\alpha|} \alpha_i)}$$

На слећој слици је графички представљена Дирихлеова расподела за три променљиве :

- $\alpha = (1, 2, 3)$

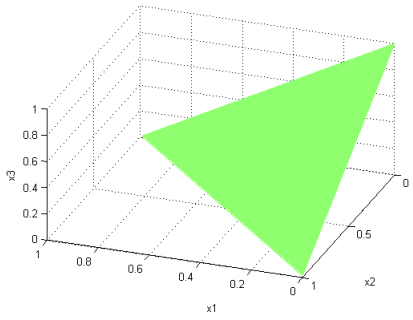


Слика 3.12: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

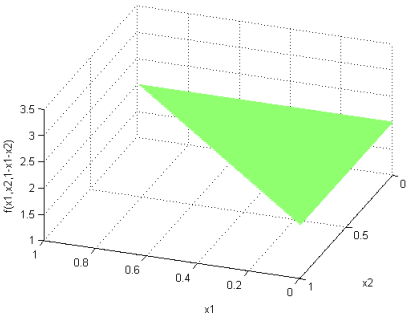


Слика 3.13: Дирихлеова расподела у три димензије

- $\alpha = (1, 1, 1)$

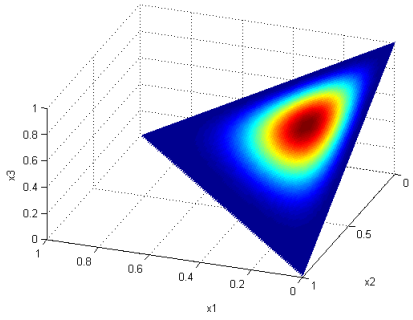


Слика 3.14: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

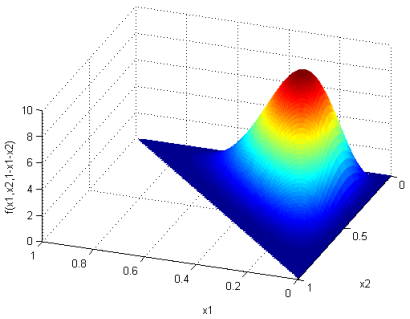


Слика 3.15: Дирихлеова расподела у три димензије

- $\alpha = (3, 3, 5)$



Слика 3.16: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу



Слика 3.17: Дирихлеова расподела у три димензије

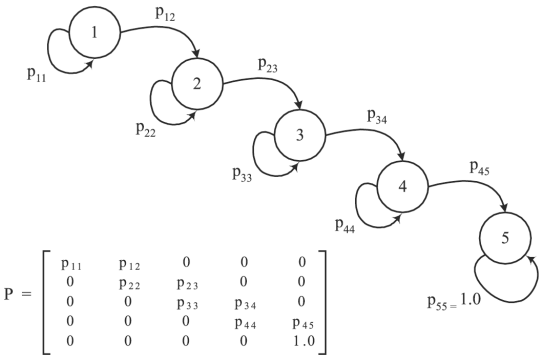
3.3 Гибсово узорковање

3.3.1 Марковљеви ланци

Марковљевим ланцима моделује се математички сиситем стања и прелаза међу тим стањима.

Дефиниција 9 *Случајан (стохастички) процес представља математички модел процеса чија је еволуција описана законима вероватноће.*
Марковљеви процеси су они случајни процеси чије будуће стање зависи само од тренутног стања. Оваква особина још се назива и одсуство памћења
Марковљеви ланци представљају посебну врсту Марковљевих процеса где се процес може налазити само у коначном броју стања.

Пример Марковљевог ланца дат је на следећој слици :



Слика 3.18: Марковљев ланац - графички пример и матрица транзиције

Систем се састоји од 5 стања. У сваком стању, са одређеним вероватноћама систем може да пређе у неко од следећих стања - конкретно да остане у тренутном стању или да пређе у једно стање ниже. Вероватноћа преласка у следеће стање зависи само од тренутног стања. Марковљеви ланци се често представљају **матрицама транзиције** при чему i-та врста у матрици садржи вероватноће преласка у свако од стање система када се систем налази у стању i. Сума свих вероватноћа у свакој врсти је 1 (систем сигурно мора да се нађе у неком стању, дакле вероватноћа да систем пређе у неко стање, могуће исто, је 1). Свака врста представља условни закон расподеле вероватноћа да систем пређе у било које стање у односу на тренутно стање (i-та врста - i-то стање). Сака колона представља маргиналну расподелу вероватноћа да се систем нађе у одређеном стању (i-та колона - i-то стање).

Систем се у једном тренутку може налазити у само једном стању. Нека се стање система карактерише случајном премнљивом X_n која у тренутку n има расподелу $\overrightarrow{s}$ и нека је укупан број стања система M. Расподела $\overrightarrow{s}$ је у ствари закон расподеле (енг. PMF) јер се ради о случајној променљивој дискретног типа - систем може бити само у једном од M могућих стања, и конкретно може се посматрати као вектор врсте, димензија $1 \times M$ где се на i - том месту налази вероватноћа да се систем у тренутку n нађе у стању i.

У следећем временском тренутку, $n+1$, систем се може наћи у било ком од M стања са различитим вероватноћама. Вероватноћа да ће се систем у тренутку $n+1$ наћи у стању j означава се са $P(X_{n+1} = j)$. Пошто ова вероватноћа зависи од стања у претходном тренутку, може се изразити на следећи начин (према формули тоталне вероватноће)

$$P(X_{n+1} = j) = \sum_{i}^{M} P(X_{n+1} = j \mid Xn = i)P(Xn = i) = *$$

$P(X_{n+1} = j \mid Xn = i)$ = вероватноћа преласка система из стања i у стање j -> $p_{i,j}$
$P(Xn = i)$ = вероватноћа да се систем у тренутку n нађе у стању i -> s_i

$$* = \sum_{i}^{M} p_{i,j} s_i$$

Дакле, верованоћа да систем у $n+1$-ом тренутку буде у стању j једнака је суми производа вероватноћа да се систем у n-том тренутку нађе у било ком стању и вероватноћа одговарајућих прелаза.

Ова сума представља j-ту колону у матрици (димензија $1 \times M$) која се добије при множењу вектора $\overrightarrow{s}$ и матрице транзиције P.

Према свему наведеном следи да је закон расподеле случајне променљиве X_{n+1} (расподела вероватноћа да се систем у $n+1$-ом тренутку налази у сваком од стања) једнак $\overrightarrow{s} \times P$.

Аналогно, у тренутку n+2, случајна променљива X_{n+2} има расподелу $\overrightarrow{s} \times P^2$, у тренутку $n+3$, случајна променљива X_{n+3} има расподелу $\overrightarrow{s} \times P^3$ итд.

Дефиниција 10 *Расподела $\overrightarrow{s}$ за коју важи :*

$$\overrightarrow{s} \times P = \overrightarrow{s}$$

*назива се **стационарна** или **равнотежна** расподела.*

$\overrightarrow{s} \times P$ представља "један корак у будућност", тј. расподелу вероватноћа да систем нађе у сваком од стања у следећем временском тренутку. Уколико расподела остаје иста, односно, вероватноће се не мењају са временом, тада се та расподела назива стационарном. Под одређеним условима везаним Марковљеве ланце, доказује се да Марковљев ланац **увек** конвергира ка својој стационарној расподели без обзира на полазно стање. Више о конвергенцији Марковљевих ланаца може се наћи у [?]. Дакле, полазне стање се може изабрати потпуно случајно а затим, уколико се дозвволи да "протекне"довољно времена, закон расподеле вероватноће да се систем нађе у свим стањима система ће конвергирати ка стационарној расподели тог ланца.

Дефиниција 11 *MCMC (енг. Markov Chain Monte Carlo) методе представљају класу алгоритама који се користе за синтетичко генерисање узорака случајних променљивих из одговрајућих расподела. Овим методама се креирају Мерковљеви ланаци који као равнотежну расподелу имају расподелу из које се узимају узорци. Једна од MCMC метода је и Гибсово узорковање (енг. Gibbs sampling)*

Гибсово узорковање

Нека је дата заједничка расподела (енг. joint distribution) $p(\mathbf{z}) = p(z_1, z_2, .., z_M)$ из које је потрбено одабрати неку вредност (енг. sample) и нека је познато почетно стање Марковљевог ланца који је потребно генерисати. Сваки корак Гибсовог узорковања почиње заменом вредности једне променљиве $z_1, z_2, .., z_M$ вредношћу која се добија из **условне расподеле** те променљиве у односу на све остале. Дакле, z_i се мења вредношћу која се узима из расподеле $p(z_i \mid z_{-i})$, где је са z_i означена i-та координата вектора z а са z_{-i} сви $z_1, z_2, .., z_M$ без z_i. Ова процедура се наставља за све променљиве по неком одређеном редоследу. При довољном броју итерација, врсности вектора z ће конвергирати ка $p(z)$.

На пример, нека је дата расподела три случајне променљиве $p(z_1, z_2, z_3)$ и нека су вредности у тренутку t : z_1^t, z_2^t, z_3^t. Нека се замена вредности променљивих врши у односу на индекс, од најмањег ка највећем. Вредност z_1^t се мења новом вредношћу z_1^{t+1} која се узима (узрокује) из расподеле

$$p(z_1 | z_2^t, z_3^t).$$

Сада се вредност z_2^t мења са вредношћу z_2^{t+1} која се узима из расподеле

$$p(z_2|z_1^{t+1}, z_3^t).$$

Дакле, одмах се користи нова вредност промењиве z_1. Коначно, за промену вредности z_2^t користи се вредност z_3^{t+1} која се добија из расподеле :

$$p(z32|z_1^{t+1}, z_2^{t+1}).$$

Овим је завршена **једна итерација** Гинбсовог узорковања. Исти процес се наставља кроз низ итерација све до одређеног броја или до неког другог услова заустављања.

Описана процедура се може уопштити и на више од три променљиве и може се представити следећим псеудокодом:

1. Initialize $\{z_i : i = 1, \ldots, M\}$
2. For $\tau = 1, \ldots, T$:
   - Sample $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - Sample $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \ldots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \ldots, z_M^{(\tau)})$.
   - $\vdots$
   - Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$.

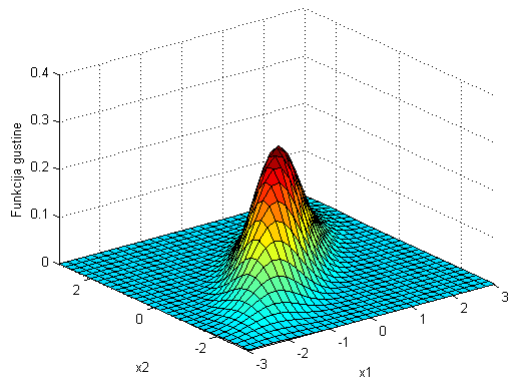
Слика 3.19: Псеудокод Гибсовог узорковања, преузето са [?]

Гибсово узорковање подразумева да су унапред познате **условне расподеле** свих променљивих и да је могуће узорковање из њих.

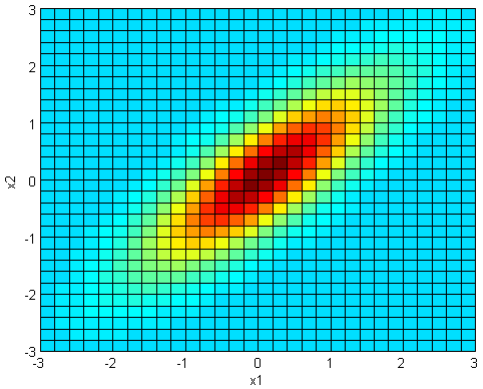
Пример : Нека је потребно узорковати вредности из дводимензионалне нормалне расподеле $\mathcal{N}(\mu, \Sigma)$ Гибсовим узорковањем при чему је

$$\mu = [\mu_1, \mu_2] = [0, 0]$$
$$\Sigma = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Графички приказ овакве дводимензионалне нормалне расподеле дат је на следећој слици (тродимензионално и пројектовано на две димензије):



Слика 3.20: Тродимензионални приказ дводимензионалне нормалне расподеле



Слика 3.21: Дводимензионлни приказ дводимензионалне нормалне расподеле

Основна претпоставка Гибсовог узорковања је да су познате условне расподеле свих променљивих и да је из њих могуће узорковати. Према [?] и [?], за условне расподеле дводимензионалне заједничке расподел важи :

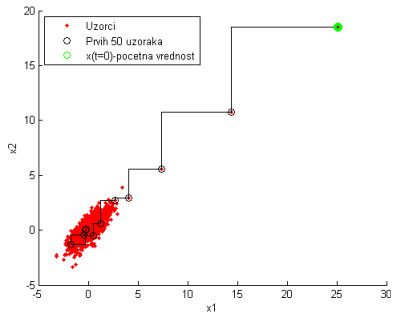
$$p(x_1 \mid x_2^{(t-1)}) = \mathcal{N}(\mu_1 + \rho_{21}(x_2^{(t-1)} - \mu_2), \sqrt{1 - \rho_{21}^2})$$

и

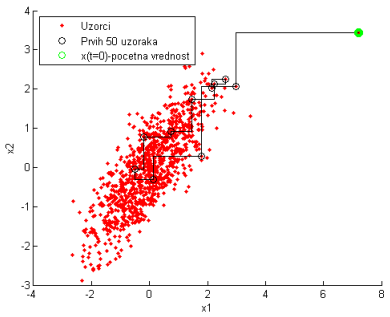
$$p(x_2 \mid x_2^{(t)}) = \mathcal{N}(\mu_2 + \rho_{12}(x_2^{(t)} - \mu_2), \sqrt{1 - \rho_{12}^2})$$

Дакле, обе условне расподеле представљају једнодимензионалну нормалну расподелу са одговарајућим параметрима. Почетне вредности променљивих се бирају случајно зато што нису од важности. Марковљев ланац ће свакако конвергирати ка дводимензионалној нормалној расподели са неведеним параметрима после одређеног броја итерација. У зависности од полазног стања, тај број итерација ће бити мањи или већи.

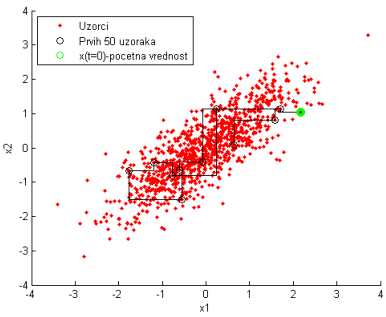
На следећем сликама су предстваљене добијене расподеле Гибсовим узорковањем за различите почетне вредности :



Слика 3.22: Почетна тачка (25.1284,18.5165)



Слика 3.23: Почетна тачка (7.2162, 3.4380)



Слика 3.24: Почетна тачка (2.1649, 1.0314)

Са слика је очигледно да се првих неколико узорака може занемарити (3.22 може се занемарити првих 7-8 узорака).

Важно је приметити да се узимање узорака увек креће по "степенастом"обрасцу. Дакле, две суседне тачке имају исту једну координату (x или y). То је зато што Гибсово узорковање у једном тренутку мења **само једну** променљиву у односу на одговарајућу вредност друге.

Конвергенција алгоритма Гибсовог узорковања ка стационарној расподели Марковљевог ланца је теоретски загарантована, али је у пракси јако тешко одредити број итерација након којих ланац почиње да конвергира. Један од начина процене конвергенције је и рачунање *log-likelihood* -а

3.4 Како ради TM алгоритам

Раније је неформално описан LDA генеративни процес. Основна претпоставка је да се сваки документ у одређеној пропорцији говори о свакој теми (има одређену расподелу над темама) као и да свакој теми све речи из корпуса припадају са различитим вероватноћама (расподела над речима). Генеративни процес се , према [?], може описати следећим псеудокодом :

1. For $k = 1...K$:
   (a) $\phi^{(k)} \sim Dirichlet(\beta)$
2. For each document $d \in \mathbf{D}$:
   (a) $\theta_d \sim Dirichlet(\alpha)$
   (b) For each word $w_i \in d$:
      i. $z_i \sim Discrete(\theta_d)$
      ii. $w_i \sim Disctete(\phi^{(z_i)})$

Слика 3.25: Генеративни процес LDA-а

при чему је : K - укупан број тема у колекцији $\phi_{(k)}$ - расподела над свим речима из колекције и представља расподелу над речима у k-тој теми θ_d - расподела над темама у документу d. z_i - тема којој припада реч w_i. α, β - **хиперпараметри** тј. параметри симетричних Дирехлеових расподела.

Описани генеративни процес резултује формирањем следеће заједничке расподеле :

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\phi \mid \beta)p(\theta \mid \alpha)p(z \mid \theta)p(w \mid \phi_z) \tag{3.16}$$

Непозанте променљиве које је потребно "открити"су z, θ и ϕ на основу (једино) познатих **речи** и њиховог присуства у сваком од докумената. Дакле, потребно је пронаћи расподеле наведених променљивих **под условом** да су познате речи и њихова распоређеност по документима тј. открити њихове постериорне расподеле. Основни проблем ТМ је **постериорно закључивање** (енг. posterior inference) односно отривање постериориних расподела латентних случајних променљивих на основу задатог скупа докумената и речи што представља решавање следеће једначине :

$$p(\theta, \phi, z \mid w, \alpha, \beta) = \frac{p(\theta, \phi, z, w, \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \tag{3.17}$$

Према [**?**], рачунање имениоца овог разломка је готово немогуће па се стога прибегава апроксимативним методама каква је и Гибсово узорковање.

Да би се применило Гибсово узорковање, потребно је познавати условне расподеле свих променљивих из чије се заједничке расподеле узоркује. Међутим, показује се да је довољно пронаћи само z јер се остале две променљиве могу преко ње израчунати и то (према [**?**]):

$$\theta_{d,z} = \frac{n(d, z) + \alpha}{\sum_{|Z|} n(d, z) + \alpha}$$

$$\phi_{z,w} = \frac{n(z, w) + \beta}{\sum_{|W|} n(z, w) + \beta}$$

Овако примењен алгоритам Гибсовог узорковања назива се још и енг. Collapsed Gibbs Sampling. Дакле, циљ је пронаћи за сваку реч, вероватноћу да припадне свакој од тема, под условом да су познате теме којима припадају остале речи у том тренутку. Формалније, ово се може записати $p(z_i \mid z_{-i}, \alpha, \beta, w)$ где z_{-i} представља доделу тема свим речима сем i-те.

Априорне расподеле коришћене у ТМ су Дирихлеове.Важна особина Дирихлеове расподеле је да је она **конјугована** са мултиномијалном расподелом. Дирихлеова расподела је расподела над параметрима мултиномијалне расподеле. Нека је на почетку претпостављено да параметри мултиномијалне расподеле припадају некој Дирихелеовој расподели - $\mathbf{p} \sim Dir(\mathbf{p}, \alpha)$. Нека је $\mathbf{x}$ узорак генерисан из мултиномијалне расподеле $Mult(\mathbf{x}; \mathbf{p})$. Тада важи да је постериорна

расподела $\mathbf{p}$ (дакле, расподела под условом да је познат узорак $\mathbf{x}$) такође **Дирихлеова расподела** са параметром $\mathbf{x}+\alpha$ тј.:

$$p(\mathbf{p} \mid \mathbf{x}, \alpha) = Dir(\mathbf{p}; \mathbf{x}+\alpha) = \frac{1}{B(\mathbf{x}+\alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} \tag{3.18}$$

Како (3.18) представља **расподелу** то важи да је :

$$1 = \int \frac{1}{B(\mathbf{x}+\alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} = \frac{1}{B(\mathbf{x}+\alpha)} \int \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} \tag{3.19}$$

Одакле следи да је :

$$\int \prod_{i=1}^{|\alpha|} p_i^{x_i+\alpha_i-1} = B(\mathbf{x}+\alpha) \tag{3.20}$$

Ова једнакост је важна за даљи опис рада TM алгоритма.

Према формули условне расподеле, важи :

$$p(z_i \mid z_{-i}, \alpha, \beta, w) = \frac{p(z_i, z_{-i}, w \mid \alpha, \beta)}{z_{-i}, w \mid \alpha, \beta)} \propto p(z_i, z_{-i}, w \mid \alpha, \beta) = p(z, w \mid \alpha, \beta) \tag{3.21}$$

$p(z, w \mid \alpha, \beta)$ се може посматрати као "маргиналан расподела"две променљиве заједничке расподеле (3.16) па важи :

$$p(z, w \mid \alpha, \beta) = \iint p(z, w, \theta, \phi \mid \alpha, \beta) d\theta d\phi = \iint p(\phi \mid \beta) p(\theta \mid \alpha) p(z \mid \theta) p(w \mid \phi_z) d\theta d\phi \tag{3.22}$$

Груписањем по заједночкој зависној променљивој, претхонда једначина се може написати :

$$p(z, w \mid \alpha, \beta) = \int p(z \mid \theta) p(\theta \mid \alpha) d\theta \int p(w \mid \phi_z) p(\phi \mid \beta) d\phi \tag{3.23}$$

Оба интерграла представљају комбинацију узорка из мултиномијалне расподеле и априорне Дирихлеове расподеле. Како је Дирихлеова расподела конјугована (conjugate prior) са мултиномијалном, у подинтегралном изразу се налази "множење"две Дирихлеове расподеле са одговарјућим параметрима.

Дакле : Пошто $p(z \mid \theta)$ има мултиномијалну дистрибуцију, важи:

$$p(z \mid \theta) = \prod_{i=1}^{D} \prod_{k=1}^{K} \theta_{d,k}^{\Omega_{d,k}} \tag{3.24}$$

, где је $\Omega_{d,k}$ број који означава колико пута је тема k додељена у документу d - број речи који у документу d припадају теми k.

Члан $p(\theta \mid \alpha)$ је из основне Дирихлеове расподеле па важи :

$$p(\theta \mid \alpha) \overset{(1)}{=} \prod_{i=1}^{D} p(\overline{\mathbf{q_d}} \mid \alpha) \overset{(2)}{=} \prod_{d-1}^{D} \frac{1}{B(\alpha)} \prod_{k=1}^{K} q_{d,k}^{\alpha_k-1} \tag{3.25}$$

где је $\overline{\mathbf{q_d}}$ расподела вероватноћа тема у документу d. Расподеле вероватноћа тема по документима су независне, па је зато могуће написати (1).Расподела тема по документу се узима из Дирихлеове расподеле па је зато могуће написати (2).

Према томе, први интеграл једнакости (3.23) се записује као :

$$\int p(z \mid \theta)p(\theta \mid \alpha)d\theta = \int \prod_{i=1}^{D}\prod_{k=1}^{K}\theta_{d,k}^{\Omega_{d,k}}\prod_{d-1}^{D}\frac{1}{B(\alpha)}\prod_{k=1}^{K}q_{d,k}^{\alpha_k-1}d\theta_d \overset{(1)}{=} \prod_{i=1}^{D}\int \frac{1}{B(\alpha)}\prod_{k=1}^{K}q_{d,k}^{\Omega_{d,k}+\alpha_k-1}d\theta_d \tag{3.26}$$

Једнакост (1) следи из чињенице да су θ_d независне расподеле па се могу интегралити посебно - правило интеграције производа Према релацији (3.20) претходна једнакост се може написати и као се :

$$\int p(z \mid \theta)p(\theta \mid \alpha)d\theta = \prod_{i=1}^{D}\frac{B(\Omega_d+\alpha)}{B(\alpha)} \tag{3.27}$$

где је са Ω означена матрица докумената и тема, $\Omega_{d,k}$ означава колико је пута тема k додељена у документу d а Ω_d је d-та врста те матрице. Елементи ове матрице могу се математички записати и овако :

$$\Omega_{d,k} = \sum_{i=1}^{N} I(d_i = d \wedge z_i = k) \tag{3.28}$$

где је N укупан број речи у корпусу(са понављањем).

Аналогно претходним извођењима, и други интеграл може да се упрости:

Члан $p(\phi \mid \beta)$ је из основе Дирхлеове расподеле па важи :

$$p(\phi \mid \beta) = \prod_{k=1}^{K} p(\phi_k \mid \beta) = \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\beta_v-1} \tag{3.29}$$

Члан $p(w \mid \phi_z)$ има мултиномијалну расподелу па важи :

$$p(w \mid \phi_z) = \prod_{i=1}^{N} p(\phi_{z_i,w_i}) = \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{k,v}^{\Psi_{k,v}} \tag{3.30}$$

где је $\psi \quad K \times V$ матрица а $\psi_{k,v}$ броји колико тема k била додељена речи v. Ова матрице може се још написати као :

$$\psi_{k,v} = \sum_{i=1}^{N} I(w_i = v \wedge z_i = k) \tag{3.31}$$

На основу (3.29) и (3.30) следи :

$$\int p(w \mid \phi_z)p(\phi \mid \beta)d\phi = \int \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k \tag{3.32}$$

Аналогно извођењу (3.26) (3.27) следи :

$$\int \prod_{k=1}^{K}\frac{1}{B(\beta)}\prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k = \prod_{k=1}^{K}(\frac{1}{B(\beta)}\int \prod_{v=1}^{V}\phi_{k,v}^{\psi_{k,v}+\beta_v-1}d\phi_k) = \prod_{k=1}^{K}\frac{B(\psi_k+\beta)}{B(\beta)} \tag{3.33}$$

Коришћењем једнакости (3.27) и (3.33), једнакост (3.23) се може записати као :

$$p(z, w \mid \alpha, \beta) = \prod_{i=1}^{D}\frac{B(\Omega_d+\alpha)}{B(\alpha)}\prod_{k=1}^{K}\frac{B(\psi_k+\beta)}{B(\beta)} \tag{3.34}$$

На основу (3.34) може се извести правило по коме ће Гибсов алгоритам узорковања мењати доделе тема речима. Дакле :

$$p(z_i = k|Z^{-i}, W, \alpha, \beta) = \frac{p(z_i = k, Z^{-i}, W \mid \alpha, \beta}{p(Z^{-i}, W \mid \alpha, \beta)} = \frac{p(Z, W \mid \alpha, \beta)}{p(Z^{-i}, W \mid \alpha, \beta)} \tag{3.35}$$

Именилац претходне једноакост се може написати преко условне вероватноће у следећем облику :

$$p(Z^{-i}, W \mid \alpha, \beta) = p(Z^{-i} \mid \alpha\beta)p(W \mid Z^{-i}, \alpha\beta) \overset{(1)}{=} \tag{3.36}$$

$$= p(Z^{-i})p(W^{-i} \mid Z^{-i})p(w_i) \propto p(Z^{-i})p(W^{-i} \mid Z^{-i}) = p(Z^{-i}, W^{-i}) \tag{3.37}$$

Једнакост (1) следи из чињенице да свако z_i зависи само од w_i. Од ове једнакости параметри α, β су изостављени због прегледности, али се подразумевају.

Форма једнакости (3.37) иста је као (3.31) па се једнакост (3.35) записује као :

$$p(z_i = k|Z^{-i}, W, \alpha, \beta) = \prod_{k=1}^{K} \frac{B(\psi_k + \beta}{B(\psi_k^{-i} + \beta} \prod_{d=1}^{D} \frac{B(\Omega_d + \alpha}{B(\Omega_d^{-i} + \alpha} \tag{3.38}$$

Коришћењем особина бета фунцкије, претходна једнакост се своди на :

$$p(z_i = k|Z^{-i}, w = v, W^{-i}, \alpha, \beta) = \frac{\psi_{k,v} + \beta_{w_i} - 1}{\left[\sum_{v=1}^{V} \psi_{k,v} + \beta_v\right] - 1} \left[\Omega_{d,k} + \alpha_k - 1\right] \tag{3.39}$$

Детаљно извођење може се наћи код [?] и [?].

3.4.1 Имплементација - псеудокод

Овде ће доћи мој псеудокод, ово је само привремено

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}, n_{k,w}$, and n_k
begin
  randomly initialize $\mathbf{z}$ and increment counters
  foreach *iteration* **do**
    for $i = 0 \to N - 1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
      for $k = 0 \to K - 1$ **do**
        $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$
      end
      $topic \leftarrow$ sample from $p(z|\cdot)$
      $z[i] \leftarrow topic$
      $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
    end
  end
  return $\mathbf{z}$, $n_{d,k}, n_{k,w}, n_k$
end

Algorithm 1: LDA Gibbs Sampling

Слика 3.26: Псеудокод - преузето са [?]

3.4.2 Закључивање расподеле тема у новом документу

На крају рада алгоритма добијају се расподеле тема по документима као и речи по темима. Поставља се питање да ли је на основу ових података могуће закључити „о чему говори“ нови документ. Односно, да ли се на основу расподела речи по темама и речи унутар новог документа може предвидети расподела тема у новом документу.

Формуле...

Пошто је доказано да је могуће закључити расподелу тема у новом документу, ова особина ће се надаље користити за изградњу решења постављеног пробелма.

Глава 4

Решење проблема применом алгоритма моделовања тема

Проналажење правог одговора на постављено питање може бити изузетно сложен проблем чак и за човека. Оно што је суштински важно за препознавање ваљаног одговора је разумевање **суштине** односно смисла питања. За разлику од машина, човек на основу знања, уме да наслути тај смисао а самим тим и да препозна адекватан одговор. Међутим, уколико би се пред човека ставило питање из области о којој он нема никаквог знања (не разуме значење речи) или је на језику који не разуме, врло је вероватно да би препознавање правог одговора било јако непоуздано. Са друге стране, немогуће је не запитати се шта заправо представља прави одговор на постављено питање. Данас је можда лакше него икада поставити питање и у кратком времену добити велики број одговора од различитих корисника. Учесници конверзације не морају нужно бити стручњаци из области које се тиче питања. Исто тако, велики број одговора, иако су наизглед адекватни, наилазе на осуду стручне популације. Дакле, потребно је извесно време, у коме долазо до комуникације међу различитим корисницима (давање оцена се може сматрати комуникацијим) да би се **закључило** шта је прави одговор. Сајтови који су служили као извор података су управо описаног карактера. Дакле, одговори на поствањено питање се оцењују од стране заинтересованих корисника и након неког времена са великом прецизношћу се може рећи који је адекватан одговор на постављено питање. Према томе, подаци који су овде разматрани јако су зависно од :

- атрактивности теме којом се баве - што је тема популарнија то ће већи број корисника бити укључен у давање и оцењивање одговора. Самим тим, може се сматрати да атрактивније теме имају поузданије одговоре
- природе питања - на нека питања се може одговорити врло кратко - (на пример где се ПМФ налази у Крагујевцу) док друга питања захтевају опширне одговоре (нпр. детаљан опис историје Крагујевца или препричано књижевно дело)

У правом одговору не морају нужно да се нађу речи из питања. Исто тако, не мора постојати законитост између дужине питања и одговора. Према томе, не постоји **алгоритам** којим се може закључити који је адекватан одговор а да се при томе не укључи додатно експертско знање.

Основна идеја овог рада је испитивање да ли се и у којој мери вештачно знање које се добија применом алгоритма моделовања тема може употребити за селектовање правог одговора без убацивања додатног експертског знања.

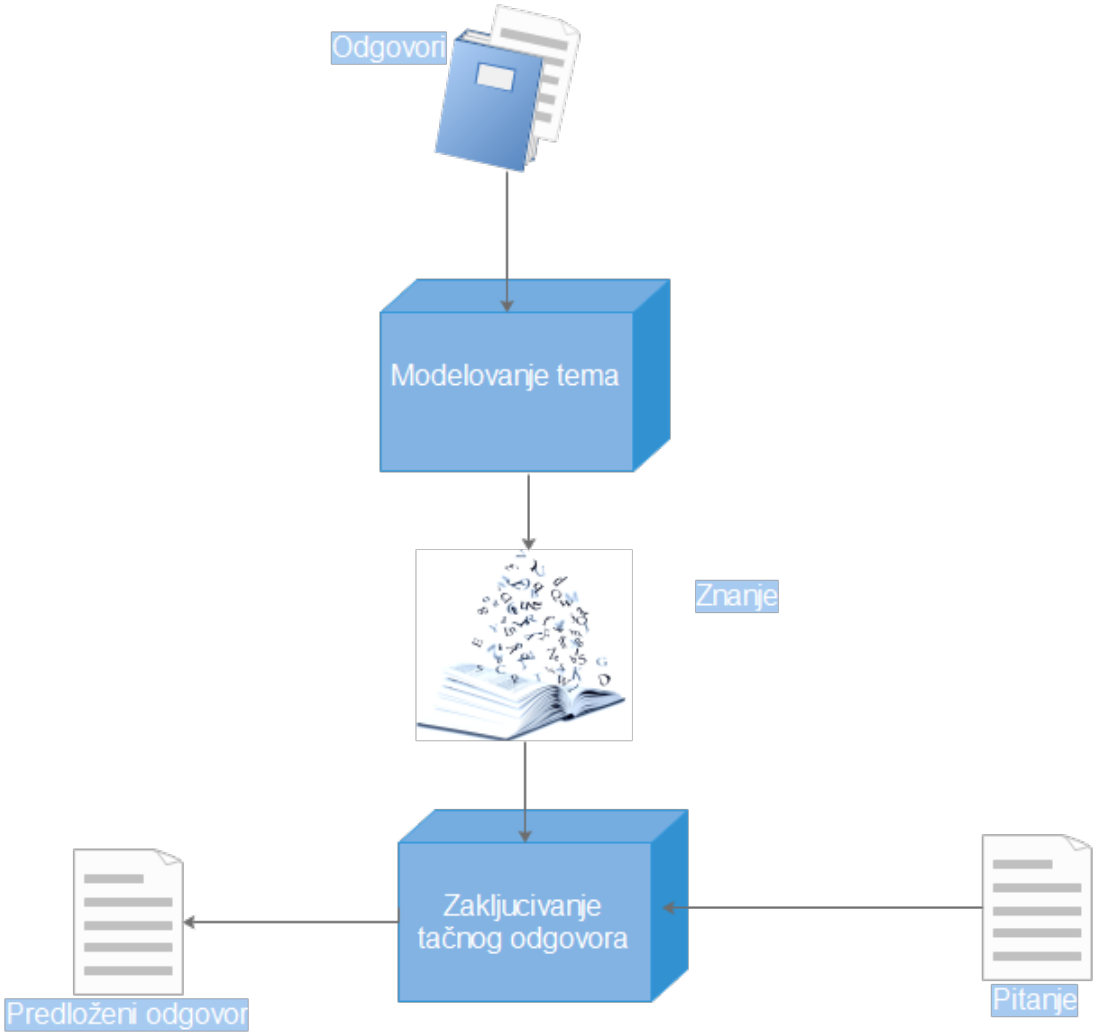
4.1 Опис решења

Идеја решења је изградња **модела тема** над свим одговорима чиме би се добио истрениран модел који поседује знање о том скупу докумената. Под знањем се подразумева расподела тема над документима као и расподела речи унутар тема. Основна предпоставка је да се при постављању питања то знање може употребити како би се селектовао тачан одговор. Даље се претпоставља да су питање и одговор највероватније из истих области (једна или више) односно да говоре о истим темама. Овде је важно напоменути да теримн **област** или **тема** више није упоредив са човековим схватањем тема или области. Обзорим да је број тема који се у истраживању користио јако велики - од 100 до 2000 - и да тема није ништа друго до скуп речи као и да у систем није укључено никакво додатно експретско знање, ово напомена је сасвим очигледна.

Селектовање правог одговора вршено је мерењем **сличности** између постављеног питања и свих одговора. Онај одговор који је **најсличнији** постављеном питању, одабира се као тачан одговор на постављено питање. Обзиром да је познато који одговор припада ком питању, рад програма је једноставно проверити и измерити.

За мерење сличности питања и одговора коришћено је неколико метода које се разликују по прецизности, брзини рада и меморијским захтевима.

Идејни ток решења може се представити дијаграмом као на слици 4.1:



Слика 4.1: Ток решења

4.2 Мерење сличности

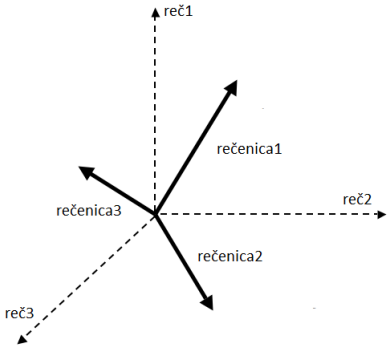
Мерење сличности је од суштинске важности за одабирање одговора на постваљено питање. Она утиче на прецизност решења, брзину извршавања, меморијске захтеве итд.

4.2.1 Косинусна сличност

Један од излаза алгоритама тема је и расподела тема по документима.Такође, за сваки нови документ, могуће је предвидети расподелу по темама. Дакле, пошто се вектор расподела по темама увек може направити, корисно их је искористити као меру сличности два документа.

Вектор расподеле по темама, дужине n може се замислити као права у n-димензионаланом простору. Дакле, вектор питања и вектор одговора могу се замислити као две праве у n-димензионаланом. Што су те две праве „ближе“ једна другој, односно што је угао између њих ближи 0, то су питање и одговор сличнији. Пошто се ради о расподелама, максимална вредност коју може да узме нека координата овако дефинисаног вектора је 1, док ј еминимална вредност 0, што значи да се обе праве налазе у првом квадранту. Дакле, максимални угао који два вектора могу да граде је 90 степени и , у смислу косинуссне мере, означава да су документи потпуно различити. Близина праве питања и праве одговора означава сличну расподелу по темама. Ако се узме у обзор полазна претпоставка да питање и одговор „говоре о“ истим темама, постаје јасно због чега се угао између ових правих може узети за меру сличности два документа.

Ради илустрације, следи један прилично упрошћен и нереалан пример. Нека је скуп свих могућих речи састављен од три речи : *реч1,реч2,реч3* и нека су дате три реченице састављене од поменутих речи : *реченица1, реченица2,реченица3.* Дате реченице могу се графички представити као праве (Слика)



Слика 4.2: Зависност просечне позиције од броја тема и броја итерација

Угао између сваке две праве представља меру сличности реченица.

Уместо мерења угла између два вектора, практичније је мерити косинус тог угла. Косинус угла који заклапају два вектора може се одредити следећом формулом :

$$cos\theta = \frac{\overrightarrow{a} \cdot \overrightarrow{b}}{\|\overrightarrow{a}\| \overrightarrow{b}\|}$$

Глава 5

Преглед резултата

Приликом израде рада развијено је софтверско окружење којим су се тестирали различити приступи решавању задатог проблема. Пре свега овде се мисли на различите мере сличности које су тестиране. Основна мера квалитета решења је **просечна позиција** тачног одговора у листи свих одговора. Што је просечна позиција нижа тј. ближа 1 то се решење сматра квалитетнијим, при чему се тежи да стандардна девијација буде што мања. Поред просечне позиције, битне карактеристике решења су брзина рада и меморијски захтеви.

Свака од тестираних мера сличности има своје параметре који су у мањој или већој мери осетљиви на одабир корака у предпроцесирању. Обзиром да се ради о алгоритму моделовања тема, у зависности од мере, биће потребан различит број тема и итерација како би се постигло оптимално решење. Због тога је било неопходно испитати за сваки меру посебно како који кораци предпроцеситрања утичу и које вредности параметара алгоритма моделовања тема су оптималне за ту меру.

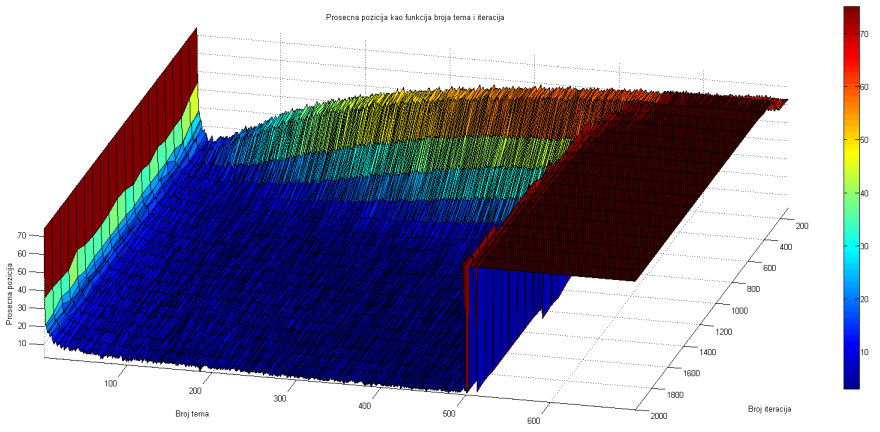
Прво тестирање решења проблема вршено је на скупу од 360 питања и 360 одговора преузетих са сајта *stackexchange.com*.

5.1 Утицај броја тема и броја итерација на просечну позицију

Обзором на то да расподела тема по документима, у мањој или већој мери, утиче на свако од тестираним мера, број тема представља битан параметар. Са малим бројем тема а на основу предложених мера, јако је тешко рангирати одговоре. Разлог томе је што ће се теме дефинисати према областима којима се баве документи. Према томе документи из исте области имаће сличне расподеле по темама а самим тим и приближно исту удаљеност од тестног документа односно питања. Због тога је неопходно да број тема буде довољно велики како би се овај проблем заобиша.

Међутим, није добро узети ни превелики број тема. У том случају сви документи ће имати сличну расподелу по темама па ће и одстојање од тестног документа бити приближно исто. Ово директно утиче на просечну позицију сводећи је на број који је једнак $[\frac{brojDokumenata}{2}]$. Овако нешто може се видети на следећем графику [1] :

[1] Ово тестирања је рађено на самом почетку и то на полазном скупу од 150 докумената. Тај скуп је подскуп скупа на коме су вршена сва остала тестирања



Слика 5.1: Зависност просечне позиције од броја тема и броја итерација

Прелаз на просечну позицију [$\frac{brojDokumenata}{2}$ (у конкретном случају 74 јер је број докумената 150) је прилично груб. Разлог томе је што нису тестиране све вредности броја итерација него свака 100-та. Како је за потребе рада било довољно уочити да се број тема не може бесконачно повећавати, ова појава даље није испитивана.

5.2 Утицај корака предпроцесирања на просечну позицију

Предпроцесирањем се **трансформише** текст улазних података како би био погоднији за обраду. Сама природа проблема намеће неке кораке предпроцесирања као неопходне. У њих спадају :

- превођење свих слова текста у мала слова - ова трансформација је неопходна како се не би реч написана почетним великим словом и иста реч написана почетним малим словом третирали као две различите речи
- уклањање HTML ознака - обзиром на порекло података, поред корисног текста, у документима се налазе и HTML ознаке које, са аспетка овог проблема, немају никаквог значаја.
- уклањање свих неалфанумеричких карактера - овом трансформацијом се уклањају знакови интерпукције и специјални знакови који нису битни за решење овог проблема
- уклањање често коришћених речи - оне не носе тематско значење, честе су у свим документима и представљју оптерећење при обради

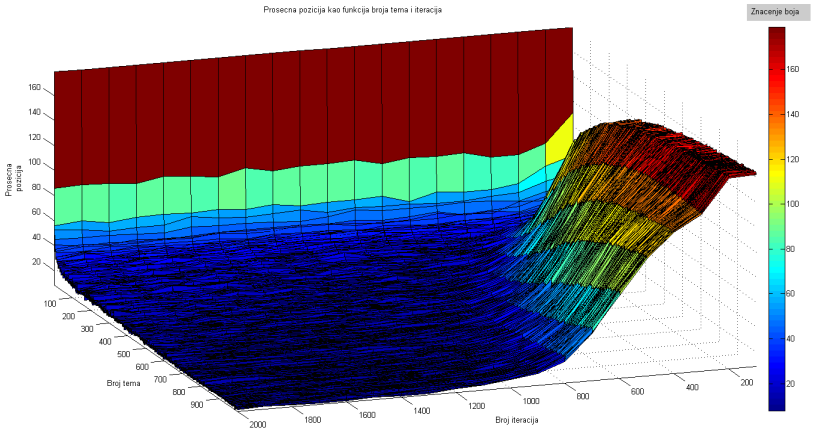
Ове трансформације су се увек укључивале, без обзира на меру сличности. Поред њих, додате су и стеминг, лемитизација и синоним трансформације у различитим комбинацијама.

5.2.1 Резултати без додатних трансформација

Улазни подаци - и питања и одговори - трансформисани преко три описана корака предпроцесирања и то у редоследу : превођење свих слова текста у мала слова, уклањање HTML ознака и уклањање често коришћених речи.

5.2.1.1 Косинусна сличност

Уколико се за меру сличности одабере **косинусна** мера, просечна позиција тачног одговора директно зависи од параметара модела тј. од броја тема и броја итерација. Та зависност се може приказати следећим графиком. (слика 5.1)



Слика 5.2: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **8** и добија се при више различитих комбинација параметара. Неке од комбинација су приказане у следећој табели

Табела 5.1: Утицај броја тема и итерација на просечну позицију тачног одговора

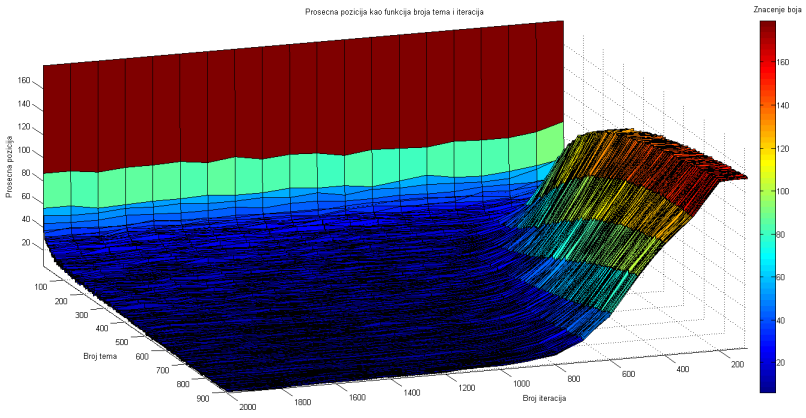
Број итерација	Број тема
1200	693
1400	756
1400	820

5.2.2 Утицај стеминга на резултат

Поред основне три трансформације, на улазне податке је примењене и стеминг модификација - уклњање наставака речи. Овим је полазни скуп података упрошћен јер је скуп различитих речи мањи. Без ове трансформације, једна иста реч у различитим родовима или временима се посматрала ралзличито што је довело до великог диверзитета у скупу речи (на пример речи енг. think и енг. thinking су посматране као различите речи без обзира на исто основно значење речи).

5.2.2.1 Косинусна сличност

Утицај стеминга на просечну позицију када се сличност мери косинусном сличношћу дат је на следећем графику(слика 5.2)



Слика 5.3: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **6** и добија се при више различитих комбинација параметара. Неке од комбинација су приказане у следећој табели

Табела 5.2: Утицај броја тема и итерација на просечну позицију тачног одговора

Број итерација	Број тема
1100	621
1200	643
1300	724

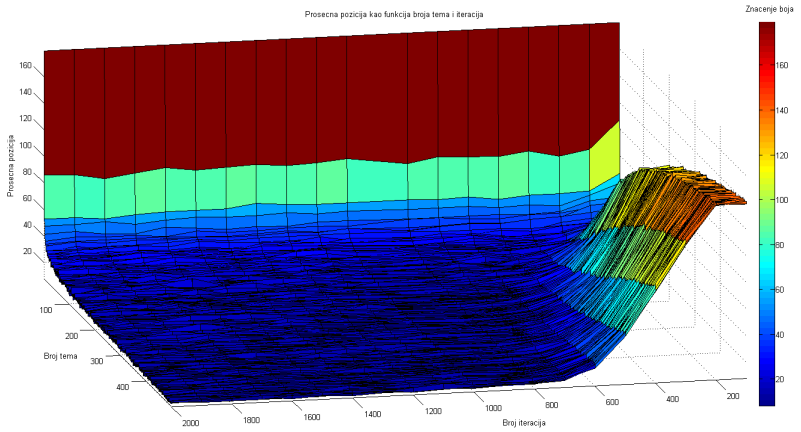
Додавање стеминга утиче на нижу просечну позицију и на упрошћавање модела. Без примене стеминга, најједноставнији модел којим се постиже оптимална просечна позиција је био комбинација 1200 - 693 док је са стемингом то 1100 - 621. Дакле, додавањем и стеминга коришћењем мањег броја тема и у мање итерација постижу се бољи резултати.

5.2.3 Утицај лемитизације на резултат

Лемитизација је процес свођења речи на **коренску** реч. Склањањем наставака речи могуће је пронаћи заједничку основу речи. Али, уколико реч мења облик у различитим лицима или временима, склањањем наставка речи, уколико уопше постоје, неће се добити иста реч. Лемитизацијом се овај проблем решева. На основу граматичких правила, препознаје се који основни облик реч и тим обликом замњује сепојављивање те речи у било којој форми.

5.2.3.1 Косинусна сличност

Утицај лемитизације на просечну позицију када се сличност мери косинусном сличношћу дат је на следећем графику(слика 5.2)



Слика 5.4: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **8** и добија се при више различитих комбинација параметара. Неке од комбинација су приказане у следећој табели

Табела 5.3: Утицај броја тема и итерација на просечну позицију тачног одговора

Број итерација	Број тема
1300	481
1300	486
1500	481
1600	477

Може се закључити да лемитизација значајно утиче на поједностављивање модела. Најмања вредност просечне позције добије се коришћењем броја тема око 470 за разлику од основног модела где тај број износи преко 600.

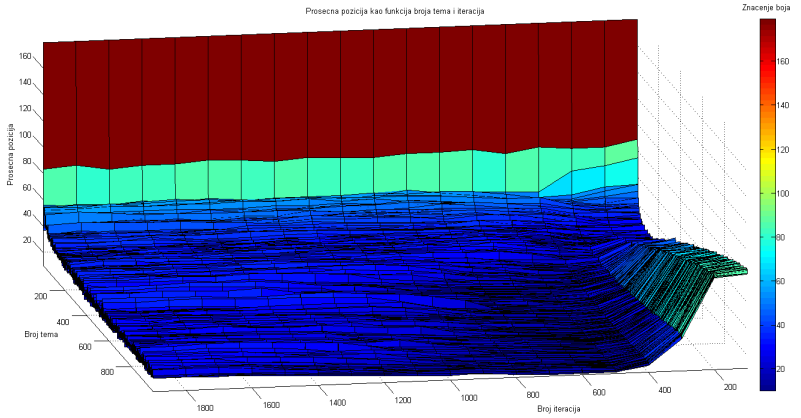
Обзиром на то да лемитизација своди на коренску реч, мањи је диверзит скупа речи. Самим тим, моделу је теже да препозна прави одговор па је због тога просечна позиција више него просечна позиција која се добија употребом стеминга.

5.2.4 Утицај додавања синонима на резултат

Додавање синонима речи има за циљ боље тематско одвајање. Наиме, уколико се свакој речи придода неколико синонима, претпоставља се да ће теметско одвајање бити једноставније. Основни разлог за увођење ове врсте трансформација била чињеница да човек поузданије закључује „о чему се ради“ у неком тексту уколико му се наведе неколико синонима за кључне речи. Обзиром да се ниједна реч не потенцира свакој речи је додат неки број синонима. Будући да додавање синонима директно утиче на перформансе система, у конкретном раду је одлучено да тај број буде 5.

5.2.4.1 Косинусна сличност

Утицај на додавања синонима на просечну позицију када се сличност мери косинусном сличношћу дат је на следећем графику(слика 5.2)



Слика 5.5: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **10** и добија се при више различитих комбинација параметара. Неке од комбинација су приказане у следећој табели

Табела 5.4: Утицај броја тема и итерација на просечну позицију тачног одговора

Број итерација	Број тема
500	674
500	675
600	773
500	991

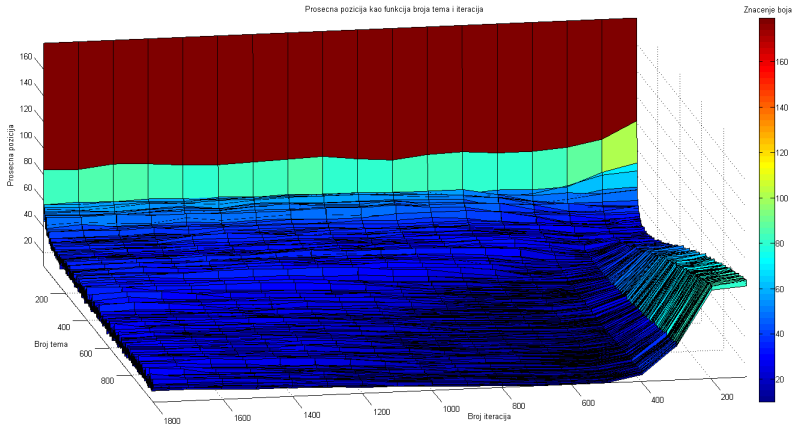
На основу облика графика може се закључити да додавање синонима утиче на екстремне вредности просечне позиције. Такође, упрошћава се модел и у мањем броју итреације достиже се минимална вредност просечне позиције. Са друге стране, пуно речи имају другачије значење у различитим контекстима а обзиром да у систем није убачено никакво додатно знање, постоји опасност од додавања неадекватних речи. То може да се одрази да немогућност прецизног одређивања тематске слике документа па и на прецизност решења.

5.2.5 Укупни резултати са синонимима и стемингом

Поред појединчане примене сваке трансформације на улазне податке, у раду су тестиране и њихове комбинације. Обзором на то да стеминг уклања наставке, велики број тако модификованих добија нерегуларни облик. Због тога није могуће ни пронаћи њихове синониме. Да би се то избегло, најпре су додавани синоними за сваку реч а затим је извршен стеминг.

5.2.5.1 Косинусна сличност

Утицај комбинације додавања синонима и стеминга на просечну позицију када се сличност мери косинусном сличношћу дат је на следећем графику(слика 5.2)



Слика 5.6: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **10** и добија се при више различитих комбинација параметара. Неке од комбинација су приказане у следећој табели

Табела 5.5: Утицај броја тема и итерација на просечну позицију тачног одговора

Број итерација	Број тема
500	555
500	690
500	712
700	847

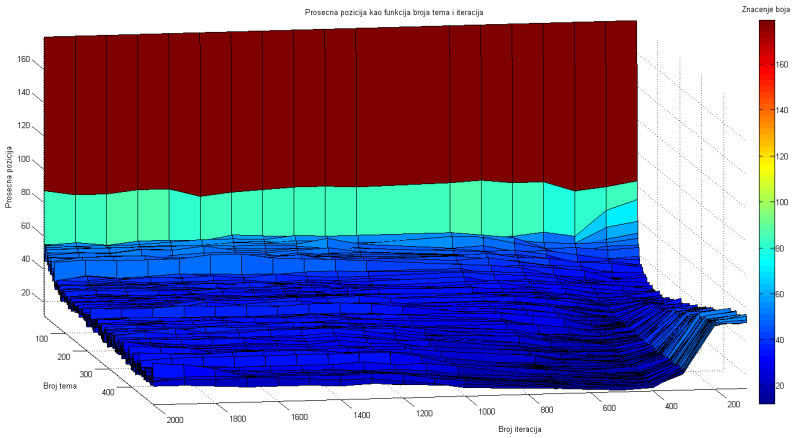
Са графика је јасно да је утицај синонима значајан. Облик површи је јако сличан облику који се добија додавањем само синонима, без примене стеминга. Дакле, додавање синонима има јачи утицај од стеминга. Односно, додавање стеминг трансформације након убацивања синонима не утиче значајно на просечну позицију.

5.2.6 Укупни резултати са синонимима и лемитизацијом

Још једна од комбинација улазних трансформација која је тестирана у овом раду је и комбинација лемитизација и синоними. Лемитизација ће свести речи на коренску, без обзира на облик те речи. Међутим, поред означавања времена и лица, различит облик речи може означавати и различито значење речи. Из тог разлога, најпре су додати синоними па је затим примењена лемитизација.

5.2.6.1 Косинусна сличност

Утицај комбинације додавања синонима и лемитизације на просечну позицију када се сличност мери косинусном сличношћу дат је на следећем графику(слика 5.2)



Слика 5.7: Зависност просечне позиције од броја тема и броја итерација

Минимална просечна позиција је **12** и добија се при тачно једној комбинацији параметара и то :

Табела 5.6: Утицај броја тема и итерација на просечну позицију тачног одговора

Број итерација	Број тема
600	483

Поново се примећује знатан утицај синонима на облик површи као и на просечну позицију. Лемитизација није значајно утицала на просечну позицију.