YASHIRIN DIRIXLE TAQSIMOTI USULI YORDAMIDA TIL KORPUSI MATNLARINI TEMATIK MODELLASHTIRISH

Elov Botir Boltayevich[1], Alayev Ruhillo Habibovich[2], Aloyev Narzillo Raxmatilloyevich[1]

[1]Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.
[2]Mirzo Ulug'bek nomidagi O'zbekiston Milliy universiteti.
[1]Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti

K E Y W O R D S

Tematik modellashtirish, tematik modellar, mashinali o'qitish, Yashirin Dirixle tadqsimoti (LDA), matnlarni klasterlash, mavzular tasnifi, NLP algortimlari.

A B S T R A C T

Tematik modellashtirish nazoratsiz ma'lumotlarda amalga oshiriladi va matnni tasniflash hamda klasterlash vazifalaridan aniq farq qiladi. Ma'lumotni qidirishni osonlashtirish va hujjatlar klasterlarini yaratishni maqsad qilgan matn tasnifi yoki klasterlashdan farqli o'laroq, tematik modellashtirish hujjatlardagi o'xshashliklarni topishni maqsad qilmaydi. Tematik modellashtirish odatda katta hajmdagi til korpusiga qo'llaniladi. Tematik modellashtirish uchta turdagi so'zlarning klasterlarini hosil qiladi – birgalikda keladigan so'zlar; so'zlarning taqsimlanishi va mavzu bo'yicha so'zlarning gistogrammasi. Tematik modellashtirishning bir nechta modellari mavjud: masalan, so'zlar sumkasi (bag-of-words), unigram modeli, generativ model. Bugungi kunda tematik modellashtirish vazifalari uchun ishlatiladigan algoritmlar sifatida Yashirin Dirixle taqsimoti (Latent Dirichlet Allocation,LDA), yashirin semantik tahlil (Latent Semantic Analysis,LSA), korrelyatsiya qilingan tematik modellashtirish (Correlated Topic Modeling) va ehtimollik yashirin semantik tahlili (Probabilistic Latent Semantic Analysis,PLSA). Ushbu maqolada LDA usuli yordamida til korpusi matnlarini tematik modellashtirish usuli tahlil qilinadi.

KIRISH

Dunyoda har kuni taxminan 2,5 kvintillion (10^{18}) bayt ma'lumot yaratiladi. Albatta, ushbu ma'lumotlarning faqat bir qismi biz bilan bog'liq bo'lgan ijtimoy tarmoq va turli axborot tizimlarida shakllantiriladi. Katta hajmdagi ma'lumiotlar sifatida ijtimoiy tarmoq xabarlari, elektron pochta xabarlari yoki fikr-mulohaza so'rovlarini miosol sifatida keltirish mumkin. Ushbu strukturlangan yoki strukturlanmagan formatlardagi ma'lumotlarni intellektual qayta ishlash uchun tematik modellashtirish deb nomlanuvchi usul va vositalardan foydalanish lozim.

Tabiiy tilni qayta ishlashda **mavzu/tema** so'zi "birga keladigan" so'zlar to'plamini anglatadi. Bir mavzu haqida biroz o'ylab ko'rganimizda, muayyan so'zlar to'plami bizning xayolimizga keladi. Misol uchun, agar biz **sport** haqida o'ylaydigan bo'lsak, unda *sportchi, futbol* va *stadion* kabi so'zlar shu mavzuga oid bo'ladi, ya'ni, **mavzular** korpus uchun *statistik ahamiyat (statistical significance)ga* ega bo'lgan takrorlanuvchi so'zlar guruhi sifatida aniqlanadi.

Statistik ahamiyatga ega bo'lish – bir nechta so'zlar bir xil hujjatlarda birga uchrashi va ularning **TF-IDF (Term frequency-Inverse document frequency)** qiymatlari o'xshash diapazonlariga ega ekanligi tushuniladi [1, 2]. Shuningdek, ushbu so'z guruhlari til korpusida muntazam ravishda doimiy tarzda paydo bo'ladi. Bu fikrlarning barchasi statistik ahamiyatga ega bo'lib, so'zlar guruhi korpus uchun muhimligini anglatadi. *O'yinlar, jamoa, xokkey, o'yin* kabi terminlarni o'z ichiga olgan so'zlar guruhi asosan **sport** mavzulariga tegishli. *Kosmos, NASA, yer, fazo, kema* va boshqalar kabi so'zlarni o'z ichiga olgan boshqa mavzular guruhi **koinot** bilan bog'liq mavzularni ifodalaydi. Endi **tematik model (topic model)** tushunchasini ko'rib chiqamiz.

Tematik model – hujjatlar to'plamida yoki korpusdagi mavzularni avtomatik ravishda aniqlaydigan tizim [3,4]. Til korpusini model vositasida o'qitish orqali shakllantirilgan tematik modeldan quyidagi maqsadlarda foydalanishimiz mumkin:

© *B. Elov, R. Alayev, N. Alayev.*

1. *ushbu mavzulardan qaysi biri yangi hujjatlarda uchrashini aniqlash;*
2. *hujjatning qaysi qismlari qanday mavzularni qamrab olishini aniqlash.*

Misol uchun, “Vikipediya” elektron resursini ko‘rib chiqamiz. Ushbu elektron resurs (web-sayt) yuz minglab mavzularni qamrab olgan millionlab hujjatlarni o‘z ichiga oladi. Shunday qilib, “Vikipediya” elektron resursida qaysi hujjatlar qaysi mavzularni qamrab olishining aniqroq xaritasini avtomatik ravishda aniqlash – muhim masala. Bu masala Vikipediyani o‘rganmoqchi bo‘lgan xalqlar/foydalanuvchilar uchun juda foydali. Yangi hosil qilingan hujjatlarni intellektual tahlil qilish natijasida yangi mavzularni ham aniqlash mumkin. Doimiy ravishda yangi hujjatlar shakllantiriladigan va yangilik bilan bog‘liq bo‘lgan ba’zi sozlamalarda (masalan, yangiliklar) **dolzarb(trend) mavzularini aniqlashga** yordam beradi.

Tematik modellashtirish

Tematik modellashtirish – bu matn ma’lumotlaridagi yashirin mavzularni aniqlashga qaratilgan avtomatik jarayon. Bu jarayon mashinali o‘qitishning nazoratsiz usuli bo‘lib, tematik modellashtirish algoritmlariga teglangan ma’lumotlar to‘plamini taqdim etish talab etilmaydi. Tematik modellashtirish usullari orqali korpus matnlaridagi mavzular model tomonidan avtomatik ravishda aniqlanadi [5, 6, 7 ,8].

Tematik modellashtirish matn korpusidagi mavzularni aniqlash va korpus matnlarini teglash jarayonini nazarda tutadi. Hujjatlarning katta to‘plamini *segmentlash*, *tushunish* va *umumlashtirish* kerak bo‘lganda, tematik modellashtirish NLP usulidan foydalanish mumkin [6, 8, 9]. Tematik modellashtirish usullari katta hajmdagi matnlarni intellektual qayta ishlash vazifalari uchun qo‘llaniladi. Ushbu yondashuv uzoq formatli tarkibni qayta ishlashga asoslangan bo‘lib, qisqa matn bilan ishlashda samarali emas. U asosan matnli ma’lumotlarga ega hujjatlarning katta to‘plamida tematik munosabatlarni topish uchun mashinali o‘rganishda qo‘llaniladi.

Mashinali o‘rganish (ML) va tabiiy tilni qayta ishlashda (NLP) tematik modellashtirish hujjatlar to‘plamida mavjud bo‘lgan mavhum “mavzular”ni aniqlashning nazoratsiz statistik usuli hisoblanadi [7, 8, 10]. U ko‘p ishlatiladigan so‘z yoki so‘z birikmalarini aniqlash uchun matnni skanerlaydi yoki tahlil qiladi va hujjatdagi ma’lumotni eng yaxshi aks ettiruvchi xulosani taqdim etish uchun ularni guruhlaydi. Hujjatda mavjud bo‘lgan yuqori darajadagi mavzularni aniqlash bilan bir qatorda, tematik modellashtirish ularga murojaat qilish darajasini ham topishini qayd etish mumkin. Ushbu ma’lumot statistik ma’lumot sifatida olinadi va mavzuning foizli taqsimotini beradi. Masalan, 60% sport haqida va 40% turizm haqida bo‘lgan maqolada, ehtimol, shahar nomlari bilan bog‘liq so‘zlarga qaraganda futbol bilan bog‘liq so‘zlar bir necha barobar ko‘p bo‘ladi.

Tematik modellashtirish hozirgidek mukammal ko‘rinish uchun ko‘plab bosqichlarni bosib o‘tdi. 1990-yilda Deervester ma’lumotni qidirish va avtomatik indeksatsiya qilish uchun Singular qiymat dekompozitsiyasini qo‘lladi va foydalanuvchi ma’lumotni so‘zlarga emas, balki tushunchaga asoslangan holda ko‘rish lozimligini ta’kidlandi [11]. U tematik modellashtirish yordamida ma’lumot olish uchun LSA va LSI usullarini taklif qildi. Korpusdan ma’lumot qidirish uchun ehtimolllik modellaridan foydalanish 1998-yilda boshlandi va generativ modeldagi so‘zlar va mavzularni bog‘laydigan PLSA yoki ehtimollik yashirin semantik tahlilga asoslangan aspekt modelini qabul qilishga olib keladi [12,13,14].

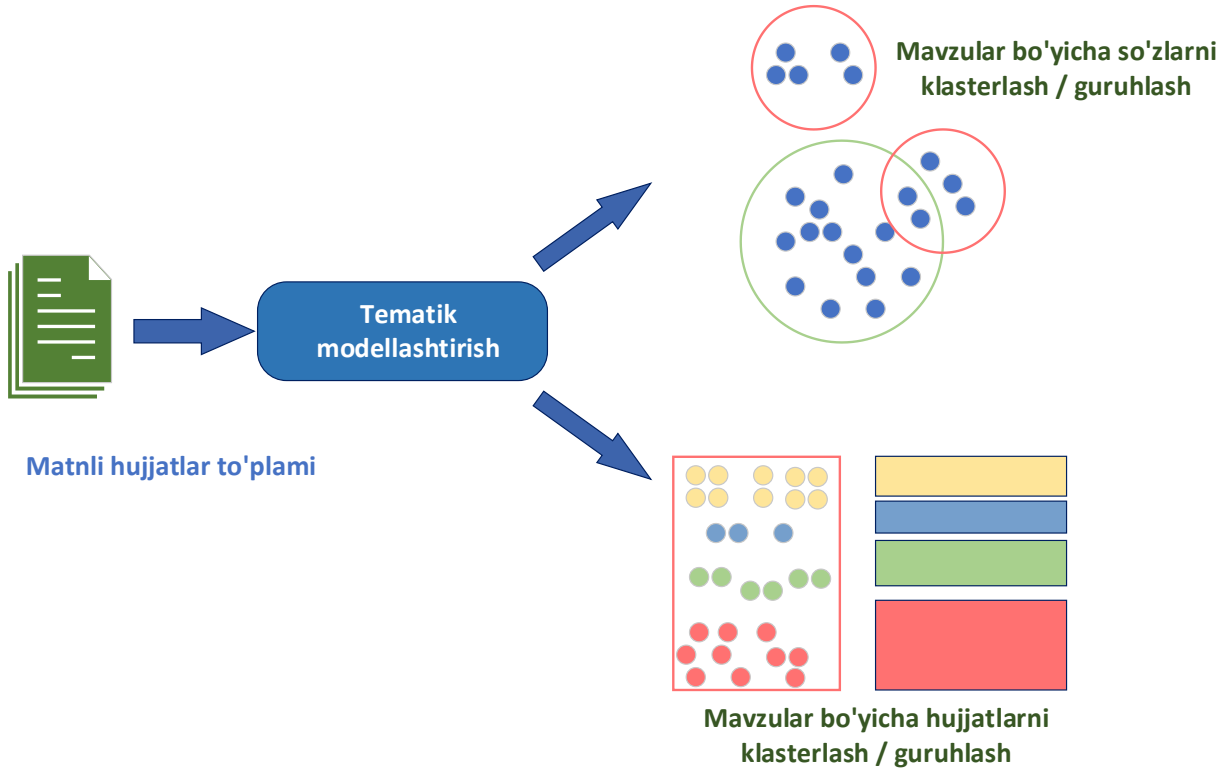
2003-yilda LDA usulining kiritilishi boshqa ko‘plab murakkab matnni intellektual tahlil qilish vazifalarida tematik modellashtirishdan foydalanish darajasini oshirdi [11,13].

2007-yilda tematik modellashtirish ijtimoiy media tarmoqlari uchun hujjatlarning ART yoki “Author Recipient” modeliga asoslangan holda qo‘llaniladi. O‘shandan beri ko‘plab o‘zgarishlar va turli xil NLP ilovalari uchun matnni intellektual tahlil qilish, tasniflash va klasterlash vazifalarini bajarish uchun ko‘plab o‘zgarishlar va yangi usullar ishlab chiqildi [15].

Tematik modellashtirish va uning usullari evolyutsiyasi dunyoning turli xil ma’lumotlarga asoslangan platformalardagi ma’lumotlarga qarashini o‘zgartirdi. Hozirgi kunda tematik modellashtirishning Hierarchal SBM for Topic

(cc) BY © B. Elov, R. Alayev, N. Alayev.

Modeling usuli ishlab chiqildi va turli NLP ilovalariga qo‘llandi [16].



1-rasm. Til korpusi matnlarini tematik modellashtirish

Tematik modellashtirish jarayonini, matndagi so‘zlarning takroriy shablon(qolip)larini aniqlash uchun matnni o‘rganish usuli sifatida ham qaraladi. Shunday qilib, tematik modellashtirish – bu hujjatlar korpusini quyidagi ikki qismga ajratish jarayoni sanaladi:

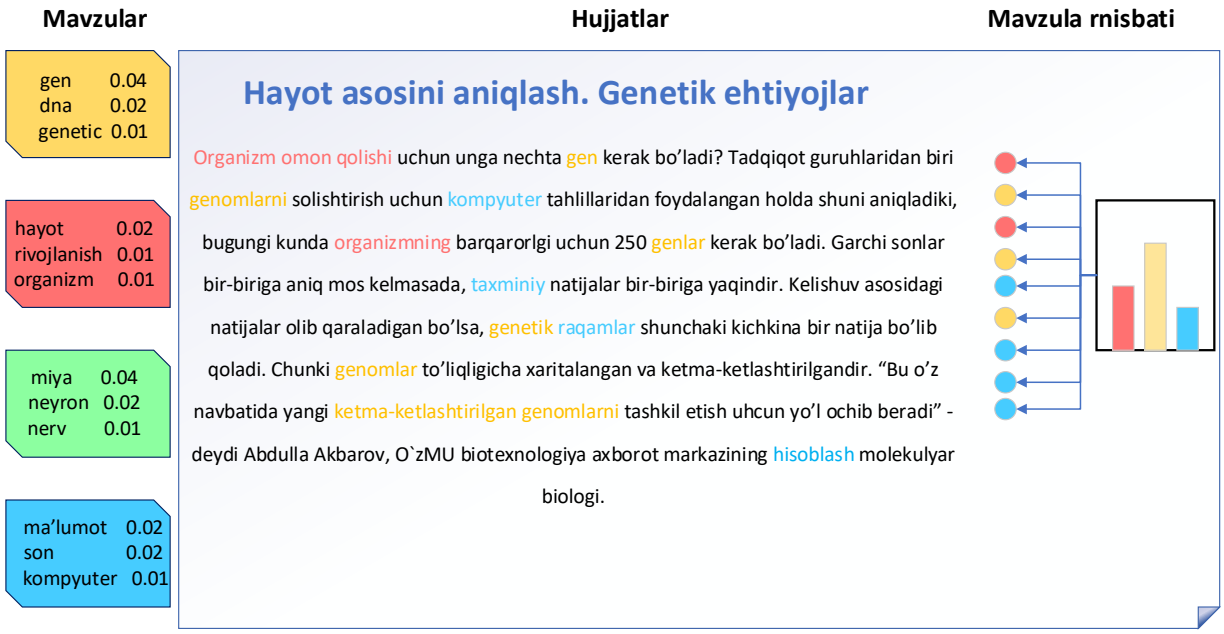
- *korpusdagi hujjatlar qamrab olgan barcha mavzularni o‘z ichiga olgan ro‘yxatni aniqlash;*
- *korpusdan bir nechta hujjatlar to‘plamini ular qamrab olgan mavzular asosida guruhlash.*

Tematik modellashtirishda har bir hujjat mavzularning statistik aralashmasidan iborat bo‘lib, bu mavzularning statistik taqsimotini anglatadi deb taxmin qilinadi. Bu esa aniqlangan barcha mavzular bo‘yicha barcha taqsimotlarni “jamlash” orqali umumiy bahoni olish mumkinligini anglatadi. Shunday qilib, tematik modellashtirishning yakuniy bosqichida korpus hujjatlarida qaysi mavzular mavjudligi va hujjatlarda ushbu mavzularning qanchalik muhim ahamiyatga ega ekanligi aniqlanadi.

Tematik modellashtirishning ahamiyati

Ma’lumki, til korpusidagi har bir hujjatni bir-birining ustiga yig‘ilgan ko‘plab mavzulardan iborat obyekt deb hisoblash mumkin. Dinamik statusga ega til korpusida har kuni katta hajmdagi ma’lumotlar qo‘shiladi. Katta hajmdagi ma’lumotlardan biz qidirayotgan narsa(axborot)ni topish murakkab vazifaga aylanadi. Shunday qilib, **katta hajmdagi ma’lumotlarni** *tartibga solish, qidirish* va *tushunish* uchun vosita va usullarni ishlab chiqish lozim.

© B. Elov, R. Alayev, N. Alayev.



2-rasm. *Korpus hujjatlarini tematik modellashtirish sxemasi*

Tematik modellashtirish ko‘p jihatdan quyidagi masalalarni hal qilishga yordam beradi:

- *hujjatlar to‘plamida mavjud bo‘lgan yashirin dolzarb shablonlarni ajratib olish;*
- *aniqlangan mavzular bo‘yicha barcha hujjatlarni izohlash/annotatsiyalash;*
- *shakllantirilgan annotatsiyalar yordamida matnlarni tartiblash, qidirish va umumlashtirish.*

Yuqorida muhokama qilinganidek, tematik modellashtirish – bu hujjatda yoki til korpusida mavjud bo‘lgan mavzulardagi so‘zlarni tanib olish. Mazkur usul orqali hujjatdan so‘zlarni ajratib olish ko‘proq vaqt talab etadi va ularni hujjatdagi mavzulardan ajratib olishdan ko‘ra ancha murakkab.

Masalan, til korpusida 1000 ta hujjat va har bir hujjatda 500 ta so‘z bor, deylik. Shunday qilib, korpusni qayta ishlashda biz 500*1000=500000 *potok (thread)ni* ko‘rib chiqamiz. Hujjat ma’lum mavzularga ajratilganida, korpus mavzular bo‘yicha teglangan bo‘lsa, unda ishlov berish potoklari soni atigi 5*500 so‘z = 2500 ta mavzudan iborat bo‘ladi. Shunday qilib, korpus hujjatlarni to‘liq qayta ishlash o‘rniga tematik modellashtirishga o‘tsak, korpusni qayta ishlashga ancha kam vaqt sarflanadi. Tematik modellashtirish hujjatlarni indekslash muammosini hal qiladi va kalit so‘zlarga mos hujjatlarni aniqlashga yordam beradi.

Tematik modellashtirish maqsadlari

Hujjatlar to‘plamidan iborat matnli korpusni ko‘rib chiqamiz. Bu hujjatlar turli xil matnlar haqida bo‘lib, tematik modellashtirish algoritmlarining asosiy maqsadi quyidagi ikkita savolga javobini aniqlashdan iborat:

1. *Hujjatlardagi eng muhim mavzular qaysi?*
2. *Har bir hujjatga qanday mavzular mos keladi?*

Tematik modellashtirish matn korpusidagi quyida keltirilgan xususiyatlarga ega bo‘lgan yashirin tuzilmani topishga harakat qiladi:

- *“mavzular”ga o‘xshaydi;*
- *to‘plamni eng yaxshi sarhisob qiladi;*
- *statistik shablonlarga asoslangan;*
- *sinonimlar, omonimlar, nomuhim so‘zlarni aniqlaydi.*

Tematik modellashtirish usuli klasterlashtirishga o‘xshaydi, ammo boshqacha “tafakkur” bilan ishlaydi:

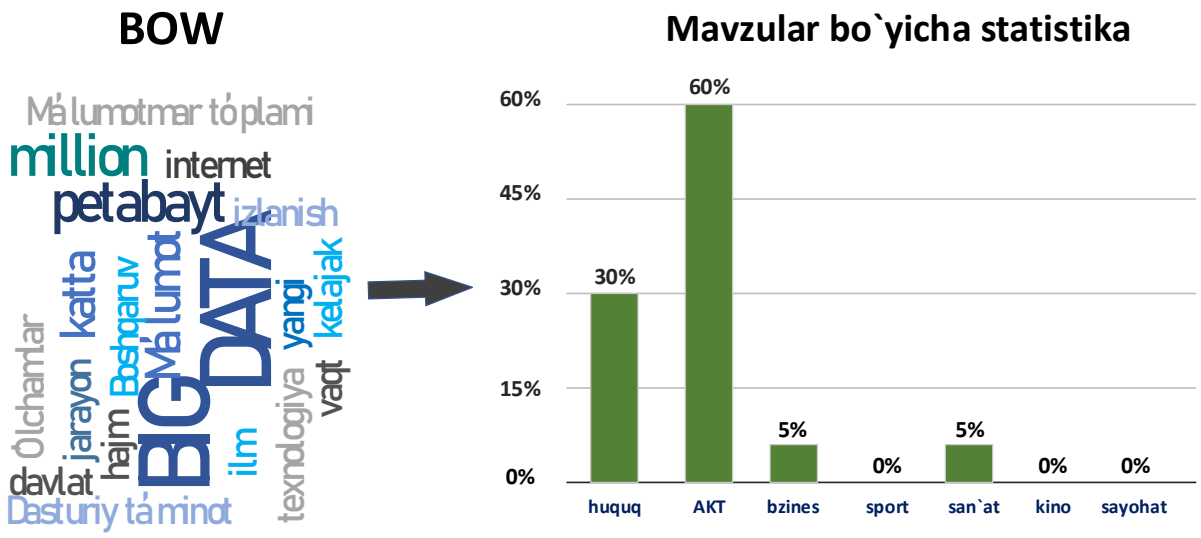
- *klasterlashda asosiy e’tibor tugun nuqtalariga/hujjatlarga qaratiladi;*
- *tematik modellashtirishda asosiy e’tibor mavzular/klasterning o‘ziga qaratiladi.*

© B. Elov, R. Alayev, N. Alayev.

Tematik modellashtirish *shovqin(noise)ni* bartaraf qilishga va matn ma’lumotlarning signalini (asosiy mavzularni) aniqlashga imkon beradi. NLP usullari yordamida korpus hujjatlaridan mavzularni ajratib olish, PCA va SVD kabi o‘lchamlarni kamaytirish usullari kabi ishlaydi [17, 18].

Tematik modellashtirish – bu korpus matnning asosiy mavzularini aniqlash uchun miqdoriy algoritmdan foydalanadi. Tematik modellashtirish PCA kabi o‘lchamlarni kamaytirish usullari bilan juda ko‘p o‘xshashliklarga ega bo‘lib, matnlardagi asosiy miqdoriy tendensiyalarni aniqlaydi. Shunday qilib, PCA usuli, misol sifatida bizga 1000 ta xususiyatdan 10 ta umumiy xususiyatga o‘tish imkonini beradi. Ushbu 10 ta umumiy xususiyat asosan mavzular sifatida ko‘rib chiqiladi. NLPda bu turdagi rolni deyarli bir xil tarzda ishlaydigan tematik modellashtirish algoritmlari bajaradi.

Misol sifatida 100000 ta xususiyatga ega bo‘lgan hujjatlarning umumiy korpusini (so‘zshakllar) 7 ta mavzuga ajratmoqchimiz. Mavzular va ular nimadan iborat bo‘lsa, qo‘limizda mavjud bo‘lgan har bir hujjatning mavzular to‘plamini **so‘zlar sumkasi (bag of words)dan** aniqlash talab qilinadi [19].



3-rasm. So‘zlar sumkasidan mavzular ro‘yxatini aniqlash

Tematik modellashtirish

Tematik modellashtirish korpusdagi so‘zlar statistikasini hisoblash va strukturlanmagan ma’lumotlardagi mavzularni aniqlash uchun o‘xshash so‘z shablonlarini guruhlashni o‘z ichiga oladi. Misol uchun, dasturiy ta’minot kompaniyasi mijozlarining dasturiy mahsulotlar haqida qanday firkrda ekanligini aniqlash lozim bo‘lsin. Ushbu vazifani tematik modellashtirish algoritmlari yordamida tahlil qilish mumkin. Shuning uchun so‘z chastotasi va so‘zlar orasidagi masofa kabi shablonlarni aniqlash orqali aniqlangan tematik model o‘xshash fikr-mulohazalarni, eng ko‘p uchraydigan so‘z va so‘z birikmalarini birlashtiradi. Ushbu ma’lumotlar yordamida har bir matn to‘plamida nima haqida gapirilayotganini tezda aniqlash mumkin. Bu yondashuv “nazoratsiz” mashinali o‘qitish usuli hisoblanib, hech qanday mashg‘ulot talab qilinmaydi.

Tematik modellashtirish usullari

Bugungi kunda tematik modellashtirishni amalga oshirishning quyidagi usullaridan foydalaniladi:

1. *Yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA)* [20,21].
2. *Manfiy bo‘lmagan matritsalarni faktorizatsiyalash (Non-negative Matrix Factorization)* [22].
3. *Yashirin semantik taqsimlash (Latent Semantic Allocation, LSA)* [23,24,25].

© B. Elov, R. Alayev, N. Alayev.

4. *Ehtimolli yashirin semantik tahlil (Probabilistic Latent Semantic Analysis, PlSA)* [13,14].
5. *Lda2Vec chuqur o‘rganish modeli (deep learning model)* [26].
6. *tBERT* [27,28].

Yuqorida keltirilgan tematik modellashtirish usullarining ba’zilarini tavsiflaymiz.

Yashirin Dirixle taqsimoti. Ehtimollar nazariyasidagi statistik noaniqliklarning barcha shakllarini tavsiflashning Bayes yondashuviga asoslanib, hujjatda ko‘rsatilgan mavzular ehtimolining to‘plamini tasvirlaydi.

Yashirin semantik tahlil. Singular qiymat dekompozitsiyasi usulidan foydalanib, hujjatlar va so‘zlarni tasniflash uchun semantik fazoda saqlashga yordam beradi.

Ehtimoliy yashirin semantik tahlil. Kutish-maksimizatsiya algoritmi bilan o‘qitilishi mumkin bo‘lib, hujjatdagi mavzu va mavzudagi so‘z ehtimolidan foydalanadi. Ushbu metodologiya so‘zlarning multinomial taqsimotiga asoslangan.

Tematik modellashtirishni aniqlash va ishlab chiqish uchun eng maqbul, ko‘p ishlatiladigan algoritm LDA yoki mavjud statistik ma’lumotlardan mavzu ehtimolini aniqlaydigan yashirin Dirixle taqsimotidir. Tematik modellashtirish metodologiyasidan foydalanishda ba’zi qiyinchiliklar mavjud. Birinchi duch keladigan muammolardan biri shundaki, tematik modellashtirish ma’lum miqdordagi mavzularni taqdim etmaydi, shuning uchun LDA yoki LSA kabi yondashuvlar haddan tashqari qayta o‘qitish, chiziqli bo‘lmaganlik va juda ko‘p umumiy so‘zlarni topish kabi muammolarni hal qilish uchun boshlang‘ich qadamlarni talab qiladi. Tematik modellashtirishda ushbu turdagi muammolarni hal qilish uchun quyida keltirilgan usullar qo‘llaniladi.

1. Matnni boshlang‘ich qayta ishlash: *lemmatizatsiya, nomuhim so‘zlari va tinish belgilarini olib tashlash*.
2. Kontekstga nisbatan kamroq uchraydigan so‘zlarni olib tashlash.
3. Mavzularni to‘plamlarda taqdim etadigan LDA usulini bajarish.
4. Sintaksisdan foydalangan holda terminlarni birlashtirish va mavzularni o‘zaro bog‘lash uchun CTM (Correlated Topic Modeling) yoki korrelyatsiya qilingan mavzularni modellashtirishni qo‘llash orqali LDAni takomillashtirish [29].

LDA usuli yordamida tematik modellashtirish

Latent Dirichlet Allocation (LDA) – til korpusi matnlarini tematik modellashtirishni amalga oshirish usullaridan biri. Bu generativ ehtimollik modeli bo‘lib, unda har bir hujjat mavzularning nisbatidan iborat deb taxmin qilinadi. Namunaviy til korpusiga tematik modellashtirish algoritmlarini qo‘llaymiz. Misol uchun, quyidagi til korpusni, ya’ni hujjatlar to‘plamini ko‘rib chiqamiz.

1-hujjat: *Men nonushta uchun yeryong‘oqli sendvich oldim.*

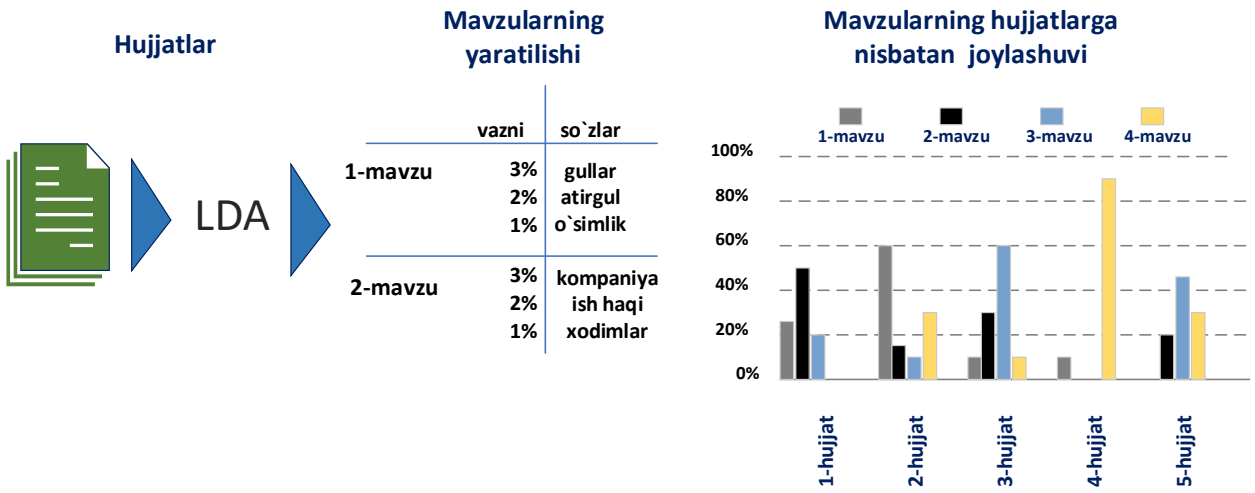
2-hujjat: *Men bodom, yeryong‘oq va yong‘oq yeyishni yaxshi ko‘raman.*

3-hujjat: *Kecha qo‘shnim kichkina it xarid qildi.*

4-hujjat: *Mushuklar va itlar o‘zaro dushmanlardir.*

5-hujjat*: Siz itingizga yeryong‘oq bermasligingiz kerak.*

© *B. Elov, R. Alayev, N. Alayev.*



4-rasm. Korpus hujjatlarini tematik modellashtirish

LDA usuli yordamida hujjatlarning har biriga mavzularni belgilash uchun quyidagi bosqichlar bajariladi:

1. har bir hujjat uchun har bir so‘zni tasodifiy ravishda K mavzular to‘plamidan tanlangan mavzuga mos qo‘yish (**K** – oldindan belgilangan mavzular soni).
2. har bir **D** hujjat uchun:

Hujjatdagi har bir W so‘zi uchun quyidagilarni hisoblash lozim:

P(t-mavzu|d-hujjati): D hujjatidagi **T** mavzusiga mos so‘zlarning nisbati;

P(w-so‘z|t-mavzu): W so‘zlari mavjud barcha hujjatlar bo‘yicha **T** mavzuga nisbati.

3. boshqa barcha so‘zlar va ularga mos mavzularni hisobga olgan holda, T` mavzusini P(T`|D)*P(W|T`) ehtimol bilan W so‘ziga qayta tayinlash.

Oxirgi 3-qadam, mavzuga mos qiymatlar o‘zgarmaydigan barqaror holatga kelgunimizcha bir necha marta takrorlanadi. Keyingi qadamda har bir hujjat uchun mavzular nisbati ushbu mavzularga mos qiymatlar asosida aniqlanadi.

LDA usulini korpus matnlariga qo‘llash

Aytaylik, bizning namunaviy matnli korpusimiz sifatida quyidagi **4** ta hujjat bor va biz ushbu hujjatlar bo‘yicha tematik modellashtirishni amalga oshirmoqchimiz.

1-hujjat: *Biz YouTubeda juda ko‘p videolarni tomosha qilamiz.*

2-hujjat: *YouTube videolari juda ham qiziqarli.*

3-hujjat: *Texnik blogni o‘qish menga narsalarni osonlik bilan tushunishga yordam beradi.*

4-hujjat: *Men YouTube videolaridan ko‘ra bloglarni afzal ko‘raman.*

LDA usuli yordamida modellashtirish bizga yuqoridagi korpusdagi mavzularni aniqlashda va har bir hujjat uchun mavzu aralashmalarini belgilashda yordam beradi. Misol tariqasida, model quyida keltirilgan natijalarni chiqarishi mumkin:

1-mavzu: *videolar* – 40%, *YouTube* – 60%;

2-mavzu: *bloglar* – 95%, *YouTube* – 5%;

U holda, 1- va 2-hujjat 100% 1-mavzuga tegishli bo‘ladi. 3-hujjat 100% 2-mavzuga tegishli bo‘ladi. 4-hujjat 80% 2-mavzuga va 20% 1-mavzuga tegishli bo‘ladi.

Hujjatlarga mavzularning bunday tayinlanishi LDA modellashtirish orqali amalga oshiriladi. Quyida ba’zi matnli ma’lumotlarga

© B. Elov, R. Alayev, N. Alayev.

LDA usulini qo‘llaymiz va Python tili vositasida natijalarni tahlil qilamiz.

```
import pandas as pd
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
rev = pd.read_csv(r"Reviews.csv")
print(rev.head())
```

I have bought several of the Vitality canned d...
Product arrived labeled as Jumbo Salted Peanut...
This is a confection that has been around a fe...
If you are looking for the secret ingredient i...
Great taffy at a great price. There was a wid...

Yuqorida keltirilgan ma’lumotlar to‘plamidagi **"Text"** ustuni uchun tematik modellashtirishni amalga oshiramiz.

Kerakli kutubxonalarni import qilish

Berilgan matnlarga boshlang‘ich qayta ishlash qadami hisoblangan lemmatizatsiya orqali tematik modellashtirish usuli samaradorligini oshirish mumkin. Shuningdek, LDA usulini bajarishdan oldin matnlardagi nomuhim so‘zlarini olib tashlash lozim. Tematik modellashtirishni amalga oshirish uchun **"Text"** ustun qiymatlarini vektorlashtirilgan shaklga aylantirish kerak. Shu sababli TfidfVectorizerni paketidan foydalanish mumkin.

```
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import
TfidfVectorizer
stop_words=set(nltk.corpus.stopwords.words('english'))
```

Matnni boshlang‘ich qayta ishlash

So‘zlarga lemmatizatsiyani qo‘llaymiz, shunda barcha hosila so‘zlarning o‘zagi qo‘llaniladi. Bundan tashqari, nomuhim so‘zlari olib tashlanadi va uzunligi 3 dan katta bo‘lgan so‘zlar ishlatiladi.

```
def clean_text(headline):
    le=WordNetLemmatizer()
    word_tokens=word_tokenize(headline)
    tokens=[le.lemmatize(w) for w in word_tokens if w
not in stop_words and len(w)>3]
    cleaned_text=" ".join(tokens)
    return cleaned_text
rev['cleaned_text']=rev['Text'].apply(clean_text)
```

Text ustuniga mos TF-IDF qiymatlarni hisoblash

Text ustunida TF-IDF vektorizatsiyasini amalga oshirish orqali tematik modellashtirishni amalga oshirishimiz mumkin bo‘lgan **hujjat termin/atama matritsasi (document term matrix, DTM)ni** shakllantiriladi. TF-IDF termin chastotasi teskari hujjat chastotasiga bog‘liq bo‘lib, bu vektorizatsiya hujjatda so‘zning paydo bo‘lish sonini so‘zni o‘z ichiga olgan hujjatlar soni bilan taqqoslaydi.

```
vect
=TfidfVectorizer(stop_words=list(stop_words),max_feat
ures=1000)
vect_text=vect.fit_transform(rev['cleaned_text'])
```

Vektorlashtirilgan matnga mos LDA qiymatlarni hisoblash

Quyida keltirilgan LDA modelidagi parametrlar, quyida ko‘rsatilganidek, **mavzular sonini**, **o‘rganish usulini** (bu algoritm hujjatlarga mavzularning topshiriqlarini yangilash usuli), **bajariladigan takrorlashlarning maksimal sonini** o‘z ichiga oladi.

```
from sklearn.decomposition import
LatentDirichletAllocation
lda_model=LatentDirichletAllocation(n_components=10
,
learning_method='online',random_state=42,max_iter=1)
lda_top=lda_model.fit_transform(vect_text)
```

Natijalarni tekshirish

© B. Elov, R. Alayev, N. Alayev.

Quyida keltirilgan dasturiy kod satrlari yordamida birinchi hujjatga tayinlangan mavzular nisbatini tekshirishimiz mumkin.

```
print("Document 0: ")
for i,topic in enumerate(lda_top[0]):
  print("Topic ",i,": ",topic*100,"%")
```

Document 0:

Topic0: 2.199225168498778 %

Topic1: 2.1985074597349286 %

Topic2: 16.96417966764352 %

Topic3: 2.1983810496807763 %

Topic4: 2.198736736277095 %

Topic5: 65.44707544380285 %

Topic6: 2.1984514162177846 %

Topic7: 2.1982341030877324 %

Topic8: 2.198764471174887 %

Topic9: 2.1984444838816577 %

Mavzularni tahlil qilish

Quyida mavzularni o‘z ichiga olgan eng yaxshi so‘zlar qaysi ekanligini tekshirib ko‘ramiz. Bu bizga ushbu mavzularning har birini belgilaydigan ko‘rinishni beradi.

```
vocab = vect.get_feature_names_out()
for i, comp in enumerate(lda_model.components_):
    vocab_comp = zip(vocab, comp)
    sorted_words = sorted(vocab_comp, key= lambda x:x[1], reverse=True)[:10]
    print("Topic "+str(i)+": ")
    for t in sorted_words:
        print(t[0],end=" ")
        print("n")
```

_____

Topic0:

chip,snack,cooky,chocolate,taste,like,peanut,butter,great, good

…

Topic9:

shipping,salt,arrived,order,price,candy,delivered,good,fast,great

Korpus matnlarini temarik modellashtirishning LDA usulidan tashqari boshqa algoritmlari ham mavjud. *Yashirin semantik indekslash (Latent Semantic Indexing, LSI), manfiy bo‘lmagan matritsalarni faktorizatsiya qilish (Non-negative matrix factorization)* kabi temarik modellashtirish usullaridan foydalanib NLPning ushbu vazifasini amalga oshirish mumkin. Bu barcha algoritmlar, masalan, LDA, hujjat termin matritsalaridan xususiyatlarni ajratib olishni va bir-biridan farq qiluvchi terminlar guruhini yaratishni o‘z ichiga oladi. Bu esa oxir-oqibat mavzularni yaratishga olib keladi. Ushbu mavzular korpusning asosiy mavzularini baholashda va matnli ma’lumotlarning katta to‘plamlarini qayta ishlashga yordam berishi mumkin.

Tematik modellashtirishni amalda qo‘llanilishi

Tematik modellashtirish til korpusidagi bir nechta mavzularni ko‘rib chiqish, ularni *tartibga solish*, *tushunish* va *umumlashtirish* imkonini beradi. Tematik modellashtirish orqali korpus ma’lumotlari bo‘ylab mavjud bo‘lgan yashirin dolzarb shablonlarni tez va oson aniqlash; keyin ma’lumotlarga asoslangan qarorlar qabul qilish uchun ushbu tushunchadan foydalanish mumkin. Misol uchun, odamlar sizning kompaniya mahsuloti haqida ijtimoiy tarmoqlarda nima deyayotganini tahlil qilish uchun mahsulotning qaysi jihatlari yoki xususiyatlari (mavzulari) ko‘p muhokama qilinayotganini aniqlashda tematik modellashtirish usullarini hissiyot tahlili bilan birlashtirish lozim. Keyin odamlarning mahsulot haqidagi tajribasi *ijobiy*, *salbiy* yoki *neytral* ekanligini bilish uchun hissiyot tahlilidan foydalanish mumkin.

Xuddi shunday, agar mijozning muammolarini tushunish uchun ularga xizmat ko‘rsatish elektron pochta xabarlarini tahlil qilish lozim bo‘lganda qanday amallarni bajarish lozim? Ushbu strukturlanmagan ma’lumotlarni biror axborot tizimiga eksport qilish, mijozlarning asosiy muammolari (muhokama qilingan mavzular) nima ekanligini ta’kidlash va tushunish uchun tematik modellashtirish vositalaridan foydalanish mumkin. Tahlil natijalariga ko‘ra,

© B. Elov, R. Alayev, N. Alayev.

tashkilotdagi xizmatlar sifatini yaxshilash, foydalanuvchilarning imkoniyatini kengaytirish uchun ma’lumotga va harakatga asoslangan strategiyani taqdim etish osonlashadi. Bu mijozlaringizning ovozini, qayerda yordam berishni tushunishda ahamiyatli. Buning ahamiyatini quyida atroflicha izohlaymiz.

Ma’lumotlarni keng miqyosida tahlil qilish. Ijtimoiy tarmoqlardagi postlardan tortib mijozlarga xizmat ko‘rsatish xatlarigacha bo‘lgan katta hajmdagi ma’lumotlar bazasini saralash o‘rniga tematik modellashtirish algoritmlari yordamida matnni intellektual tahlil qilish jarayonini avtomatlashtirish mumkin. Bu tez, aniq va kengaytiriladigan, ya’ni siz xohlagancha ma’lumotni qayta ishlash samaradorligini baholashingiz mumkin.

Real rejimda tahlil qilish. Tematik modellashtirishni NLPning boshqa vazifalari, masalan, hissiyotlarni tahlil qilish bilan birlashtirib, foydalanuvchi tajribasi haqida to‘liq tasavvurga ega bo‘lish uchun teginish nuqtalari bo‘ylab bo‘shliqlarni yopish mumkin. Eng muhimi, butun jarayon avtomatlashtirilganligi sababli, siz xohlagan vaqtda va qayerda bo‘lmasin, tushunchani harakatga aylantirishingiz mumkin.

Ma’lumotlarni izchil tushunish. Mavjud ma’lumotlardan keng miqyosda ma’no chiqarish mumkin bo‘lsa, so‘rovlardan tortib mijozlar fikrigacha bo‘lgan har bir aloqa nuqtasidagi o‘zaro ta’sirlarni tushunish va bu tushunchadan yaxshiroq tajriba yaratish uchun foydalanish mumkin.

Demak, tematik modellashtirish tahlili sizni real vaqt rejimida tegishli ma’lumotlarni olish va undan ajoyib tajriba yaratish uchun foydalanish uchun katta hajmdagi matnli ma’lumotlarini tahlil qilish uchun ajoyib shaklga keltiradi.

Xulosa

Nazorat qilinmaydigan mashinali o‘rganish vazifalarida qo‘llaniladigan temarik modellashtirish teglash sifatida ko‘rib chiqiladi va, birinchi navbatda, til korpusidan zarur ma’lumotlarni olish uchun ishlatiladi hamda so‘rovlarning bajarilish samaradorligining oshishiga yordam beradi. Temarik modellashtirish qidiruv tizimlarida mavzular bo‘yicha foydalanuvchi qiziqishlarini xaritalashda keng qo‘llaniladi. Bugungi kunda temarik modellashtirish usullari: *hujjatlarni tasniflash, toifalarga ajratish, umumlashtirish* kabi NLP vazifalarini hal qilishda qo‘llanilmoqda. *Genetika, ijtimoiy media* kabi qator sohalarda AI metodologiyalari tematik modellashtirish bilan bog‘liq. Shuningdek, temarik modellashtirish usullari ijtimoiy tarmoqlardagi foydalanuvchilarning his-tuyg‘ularini tahlil qilish imkonini beradi.

Matnni intellektual qayta ishlash, matnni tasniflash, mashinali o‘rganish, ma’lumot qidirish va tavsiya tizimlari kabi NLP vazifalarini hal qilishda nazoratsiz va yarim nazorat qilinadigan yondashuvlar bilan birgalikda tematik modellashtirish usullarini qo‘llash mumkin. Axborotni qidirish (Information Retrieval, IR) kabi NLP vazifasini hal qilshda tematik modellashtirish muhim bosqich hisoblanadi. Matematik jihatdan, IR ilovasida ma’lumotlarni qidirish quyidagilarni o‘z ichiga oladi – hujjatlarni taqdim etish, so‘rovlar, ramka va reyting tizimi. Qo‘shimcha iqtibos keltirish uchun IR foydalanuvchi so‘roviga tegishli ma’lumot bazasini taqdim etish uchun Google, Bing kabi qidiruv tizimlari tomonidan qo‘llaniladi. Tematik modellashtirish, odatda, katta hajmli matn tarkibiga ega bo‘lgan ma’lumotlar bazasida matn tasnifini ta’minlash uchun ham qo‘llaniladi. Genomika uchun ishlatiladigan qidiruv tizimlari foydalanuvchi so‘rovlari bo‘yicha tegishli ma’lumotlarni to‘plash va taqdim etishda tematik modellashtirishdan foydalanadi.Tematik modellashtirishni qo‘llash oddiy ko‘rinadi, ammo ma’lumotni tartiblash va taqdim etish uchun qo‘llaniladigan metodologiyalarni tadqiq qilish muhim ahamiyatda ega. Ushbu maqolada namunaviy korpus matnlarini LDA usulidan foydalanib tematik modellashtirish masalasi ko‘rib chiqildi. Biroq samarali natijaga erishish uchun ushbu usulni katta hajmdagi til korpusiga qo‘llash tavsiya etiladi.

Foydalanilgan adabiyotlar

1. B.Elov, Z.Xusainova, N.Xudayberganov. O‘zbek tili korpusi matnlari uchun TF-IDF statistik ko‘rsatkichni hisoblash. *SCIENCE AND INNOVATION INTERNATIONAL SCIENTIFIC JOURNAL VOLUME 1 ISSUE 8 UIF-2022: 8.2 | ISSN: 2181-3337*

© *B. Elov, R. Alayev, N. Alayev.*

2. Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, *9*(1). https://doi.org/10.1186/s13673-019-0192-7
3. Gao, Q., Huang, X., Dong, K., Liang, Z., & Wu, J. (2022). Semantic-enhanced topic evolution analysis: a combination of the dynamic topic model and word2vec. *Scientometrics*, *127*(3). https://doi.org/10.1007/s11192-022-04275-z
4. Zou, X., Zhu, Y., Feng, J., Lu, J., & Li, X. (2019). A novel hierarchical topic model for horizontal topic expansion with observed label information. *IEEE Access*, *7*. https://doi.org/10.1109/ACCESS.2019.2960468
5. Korencic, D., Ristov, S., Repar, J., & Snajder, J. (2021). A Topic Coverage Approach to Evaluation of Topic Models. *IEEE Access*, *9*. https://doi.org/10.1109/ACCESS.2021.3109425
6. Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. *ACM International Conference Proceeding Series*, *382*. https://doi.org/10.1145/1553374.1553515
7. Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference*.
8. Waal, A. de, & Barnard, E. (2008). Evaluating topic models with stability. *Annual Symposium of the Pattern Recognition Association of South Africa*.
9. Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*.
10. Elov B., Aloyev N., Yuldashev A. SVD va NMF metodlari orqali tematik modellashtirish // *Труды XI Международной конференции «Компьютерная обработка тюркских языков» «TURKLANG 2023»*. Бухара, 20-22 октября 2023 г.
11. Deerwester, S. (1988). Improving Information Retrieval with Latent Semantic Indexing. *In Proceedings of the 51st ASIS Annual Meeting (ASIS \'88), Vol. 25 (October 1988)*, *25*.
12. Qiang, J., Qian, Z., Li, Y., Yuan, Y., & Wu, X. (2022). Short Text Topic Modeling Techniques, Applications, and Performance: A Survey. In *IEEE Transactions on Knowledge and Data Engineering* (Vol. 34, Issue 3). https://doi.org/10.1109/TKDE.2020.2992485
13. Zhuang, F., Karypis, G., Ning, X., He, Q., & Shi, Z. (2012). Multi-view learning via probabilistic latent semantic analysis. *Information Sciences*, *199*. https://doi.org/10.1016/j.ins.2012.02.058
14. Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1–2).
15. https://doi.org/10.1023/A:1007617005950
16. Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., & Steyvers, M. (2010). Learning author-topic models from text corpora. *ACM Transactions on Information Systems*, *28*(1).
17. https://doi.org/10.1145/1658377.1658381
18. Rehman, A. U., Khan, A. H., Aftab, M., Rehman, Z., & Shah, M. A. (2019). Hierarchical topic modeling for Urdu text articles. *ICAC 2019 - 2019 25th IEEE International Conference on Automation and Computing*. https://doi.org/10.23919/IConAC.2019.8895047
19. Ogunleye, B., Maswera, T., Hirsch, L., Gaudoin, J., & Brunsdon, T. (2023). Comparison of Topic Modelling Approaches in the Banking Context. *Applied Sciences (Switzerland)*, *13*(2). https://doi.org/10.3390/app13020797
20. Qiu, J., Wang, H., Lu, J., Zhang, B., & Du, K.-L. (2012). Neural Network Implementations for PCA and Its Extensions. *ISRN Artificial Intelligence*, *2012*. https://doi.org/10.5402/2012/847305
21. B.Elov, Z.Xusainova, N.Xudayberganov (2022). Tabiiy tilni qayta ishlashda Bag of

© *B. Elov, R. Alayev, N. Alayev.*

Words algoritmidan foydalanish. *O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4).*
22. http://aphil.tsuull.uz/index.php/language-and-culture/article/download/32/29
23. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, *78*(11). https://doi.org/10.1007/s11042-018-6894-4
24. Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*.
25. Wang, J., & Zhang, X. L. (2023). Deep NMF topic modeling. *Neurocomputing*, *515*. https://doi.org/10.1016/j.neucom.2022.10.002
26. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6). https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9
27. Tao, R., Wei, Y., & Yang, T. (2021). Metaphor Analysis Method Based on Latent Semantic Analysis. *Journal of Donghua University (English Edition)*, *38*(1). https://doi.org/10.19884/j.1672-5220.202010087
28. Qi, Q., Hessen, D. J., Deoskar, T., & van der Heijden, P. G. M. (2023). A comparison of latent semantic analysis and correspondence analysis of document-term matrices. *Natural Language Engineering*, *8*(10). https://doi.org/10.1017/S1351324923000244
29. Mishra, P. (2020). A Comparative Study for Sentiment Analysis: LDA and LDA2Vec. *International Journal of Emerging Trends in Engineering Research*, *8*(8). https://doi.org/10.30534/ijeter/2020/06882020
30. Peinelt, N., Nguyen, D., & Liakata, M. (2020). tBERT: Topic models and BERT joining forces for semantic similarity detection. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/2020.acl-main.630
31. Liu, Z., Zhao, K., & Cheng, J. (2023). TBERT: Dynamic BERT Inference with Top-k Based Predictors. *Proceedings - Design, Automation and Test in Europe, DATE, 2023-April*.
32. https://doi.org/10.23919/DATE56975.2023.10136977
33. He, J., Hu, Z., Berg-Kirkpatrick, T., Huang, Y., & Xing, E. P. (2017). Efficient correlated topic modeling with topic embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F129685*. https://doi.org/10.1145/3097983.3098074

© *B. Elov, R. Alayev, N. Alayev.*