TEMATIK MODELLASHTIRISHNING ZAMONAVIY USULLARI

Elov Botir Boltayevich[1], Alayev Ruhillo Habibovich[2], Aloyev Narzillo Raxmatilloyevich[3],
[1]Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti, dotsent, t.f.f.d.
[2]Mirzo Ulug'bek nomidagi O'zbekiston Milliy universiteti, t.f.f.d.
[3]Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch doktoranti.
E-pochta: vip.alayev@gmail.com

K E Y W O R D S

Tematik modellashtirish, LDA, LSA, PLSA, lda2vec, tBERT, nazoratsiz mashinali o`qitish usullari, til korpusi.

A B S T R A C T

Tematik modellashtirish – tabiiy tilni qayta ishlash (Natural Language Processing, NLP) vazifalarining muhim komponenti bo`lib, turli sohalarda ishlatilishi mumkin. Tematik modellarida ikkita asosiy muammo mavjud: natijani bashorat qilish uchun ishlatilishi mumkin bo'lgan mavzularni aniqlash va allaqachon topilgan mavzularni tushunishni osonlashtirish. Ushbu maqolada, tematik modellashtirishning zamonaviy usullari hisoblangan manfiy bo'lmagan matritsa faktorizatsiyasi (Non-negative Matrix Factorization), yashirin semantik taqsimot (Latent Semantic Allocation, LSA), ehtimoliy yashirin semantik tahlil (Probabilistic Latent Semantic Analysis,PLSA), lda2vec chuqur o'rganish modeli va tBERT haqida fikr-mulohaza yuritiladi. 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. Ushbu maqolada o`zbek tili korpusidagi namunaviy matnlarga LDA usulini qo`llash orqali olingan natijalar keltiriladi.

Kirish

Tematik modellashtirish – bu hujjatlar to'plamidagi yashirin mavhum mavzularni avtomatik ravishda aniqlash uchun statistik va mashinali o'rganish modellaridan foydalanadigan matnni intellektual qayta ishlash usullari to'plami. Tematik modellashtirish, shuningdek, hujjatlardagi so'z va so`z birikmalari shablon (qolip)larini aniqlashga va hujjatlar to'plamini samarali tushunishga yordam beradigan so'z guruhlari va shunga o'xshash birikmalarni avtomatik ravishda klasterlash qobiliyatiga ega bo'lgan nazoratsiz mashinali o`qitish usllari to'plamidir [1,2,3].

Ijtimoiy tarmoqlar yoki turli axborot tizimlarida yaratilgan katta hajmdagi matnlarda qanday ma`lumotlar muhimligini tushunish uchun butun matnni ko'rib chiqish va tahlil qilish juda ko`p vaqt va resurni talab qiladi [2,4]. Bunday hollarda NLP algoritmlari va xususan tematik modellashtirish usullari asosiy matnning xulosasini chiqarish va matndan muhim kontekstlarni aniqlash uchun foydalidir.

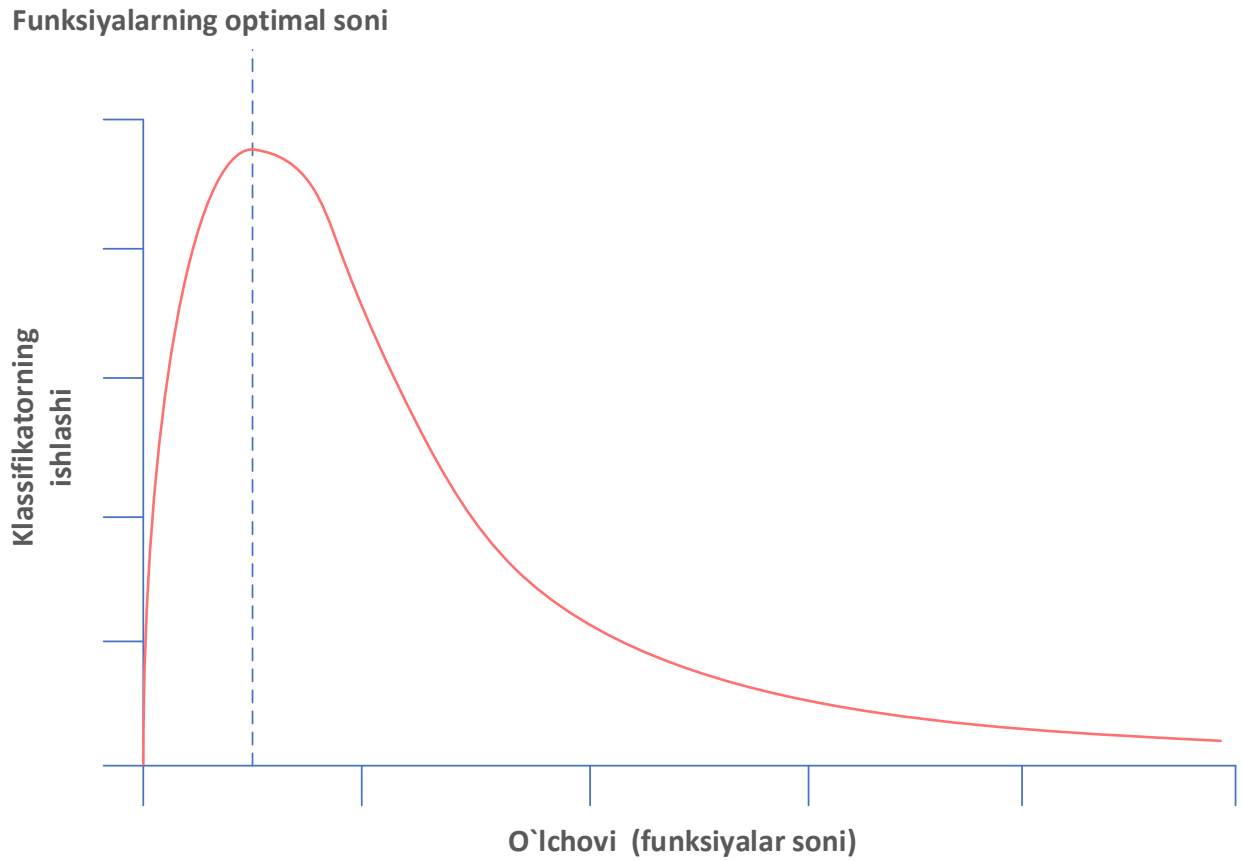
NLPda tematik modellashtirishning asosiy maqsadi o`zaro bog'liq bo'lgan so'zlarning birikmasi sifatida ifodalangan so'zlar klasteri bo'lgan mavzularni aniqlashdir. Har bir mavzu bir yoki bir nechta hujjatlarga tegishli bo'lganligi sababli, o'z navbatida har bir hujjatni bir yoki bir nechta mavzularning kombinatsiyasi sifatida ifodalash mumkin [2,3,5].

O'lchov muammosi (Curse of Dimensionality,CoD)

NLP modeliga qo'shimcha funktsiyalar qo'shilisi sababli, model samaradorligini oshirish uchun o'rgatish uchun zarur bo'lgan ma'lumotlar miqdori eksponent ravishda oshadi (1-rasm). O'lchovlar NLP va ML modellari kontekstidagi xususiyatlar hisoblanadi [4,6,7]. O`lchovlar, ayniqsa, mashinali o'rgatish algoritmlari uchun juda muhim bo`lib, bunda modelning aniqligi

© Elov B.B., Alayev R.H., Aloyev N.R.

dastlab xususiyatlar soni ortishi bilan ortadi, lekin o'lchovlilik oshishi tufayli ma'lumotlarni doimiy ushlab turish bilan o'lchovlilik oshgani sayin yomonlasha boshlaydi.



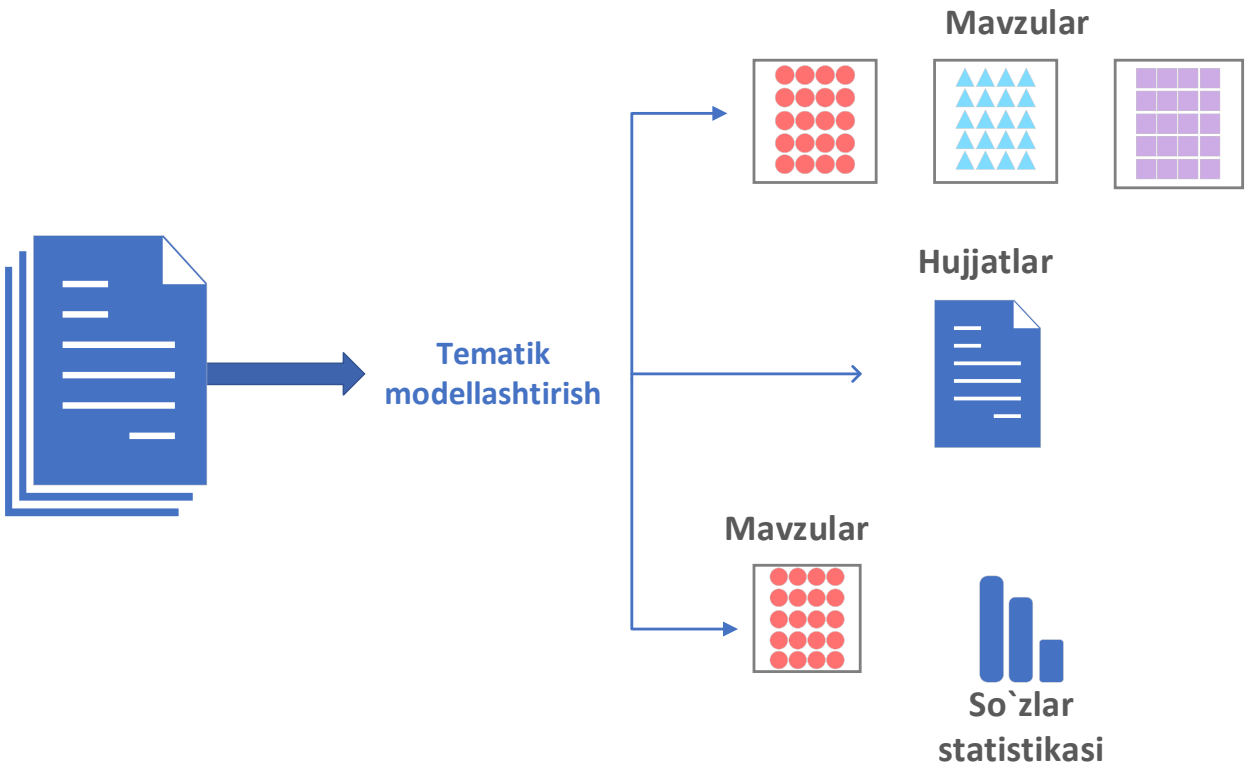
1-rasm. *NLPda mashinali o`qitishda o'lchov muammosi*

NLP modellariga asoslangan tematik modellashtirishda o'lchov muammosi muhimdir. Chunki mashinali o`qtitish modellari orqali juda katta hajmdagi til korpusiga mos juda ko'p sondagi xususiyatlar aniqlanadi. Shuning uchun katta o'lchamli modellar bilan ishlashda xususiyatlarni tanlash yoki o'lchovni kamaytirish usullarini ishlab chiqish lozim [7].

NLPda tematik modellashtirish

Hujjatlar to'plamini tahlil chiqish jarayonida, har bir hujjat bir nechta fikrlar yoki g'oyalarning kombinatsiyasi bo'lishi mumkin. Bu holda, ushbu fikrga tegishli so'zlar hujjatlar bo'ylab turli nisbatlarda aniqlanadi. Shu sababli, har bir hujjatdagi so'zlarning statistikasini aniqlash va o'xshash so'zlarning klasterlari bo'lgan mavzularni yaratish mumkin [8]. Tematik modellar bizga hujjatlarning butun matnida o'xshash ma'noga ega so'zlar birikmasi va har bir berilgan hujjatda mavzular birikmasi sifatida yashiringan potentsial mavzularni aniqlashga yordam beradi (2-rasm).

© Elov B.B., Alayev R.H., Aloyev N.R.

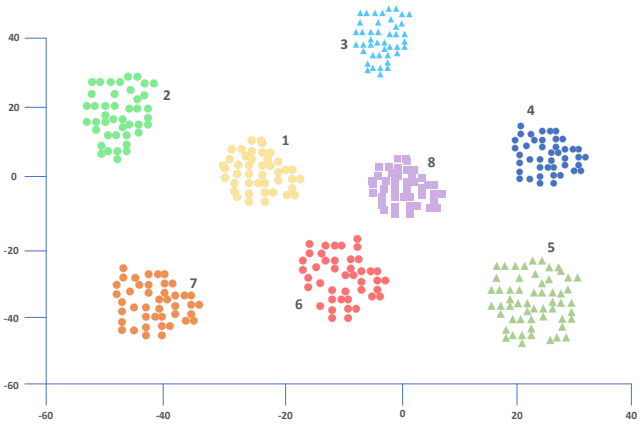


2-rasm. *Tematik modellashtirish jarayoni*

Asosiy komponentlar tahlili (Principal Component Analysis, PCA)

Asosiy komponentlar tahlili statistik usul bo'lib, asosiy g'oya ma'lumotlar to'plamida mavjud bo'lgan o'zgarishlarni imkon qadar ko'proq saqlab qolgan holda ko'p sonli o'zaro bog'liq o'zgaruvchilardan iborat ma'lumotlar to'plamining *o'lchamini kamaytirishdir*. PCA algoritmining ishlashi tamoyili quyidagicha [8,9,10]:

- *Ma'lumotlar to'plamidagi ustunlarni **asosiy komponentlar (principal components, PCs)** deb nomlangan yangi o'zgaruvchilar to'plamiga aylantirish lozim.*
- *Asosiy komponentlar o'zaro bog'liq emas va ularning xarakterli qiymatlari asosida tartibga solinishi mumkin. Shuning uchun barcha asl o'zgaruvchilarda mavjud bo'lgan o'zgarishlarning ko'p qismini saqlab qolgan holda birinchi navbatda bir nechta asosiy komponentlarni tanlash mumkin.*



3-rasm. *Asosiy komponentlar tahlili jarayoni*

Yuqoridagi 3-rasmdagi ranglar ma'lumotlar to'plamidagi maqsadli o'zgaruvchilar / guruhlar / klasterlar teglarini, koordinata o'qlari esa PCA usuli yordamida olingan dastlabki ikkita asosiy komponentni ifodalaydi.

PCA usuli quydagi afzalliklarga ega:

- *PCA – bu mashinali o'qitishda boshlangi`ch ishlov berish uchun samarali va keng qo'llaniladigan usul;*
- *PCA usuli modelni o'qitish va baholashga sarfalanadigan vaqtini tejashfa yordam beradi,*

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

chunki u kiritilgan o'zgaruvchilar sonini sezilarli darajada kamaytiradi;

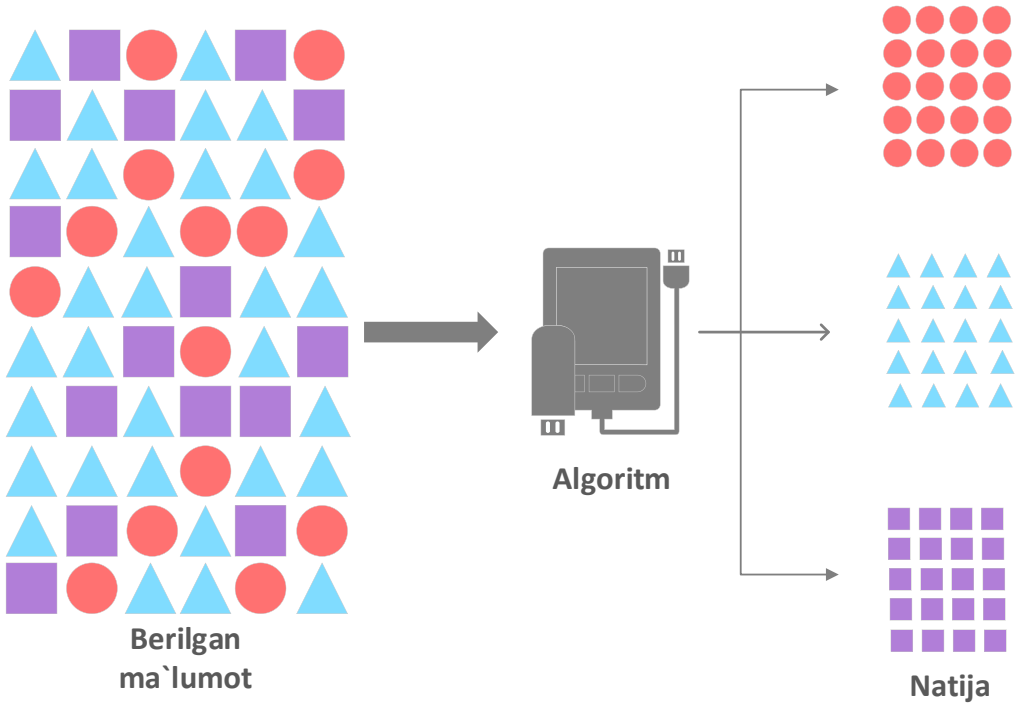
- *PCA shuningdek, o'lchamlilik muammosini hal qilishda modellarning aniqligini oshirishga yordam beradi va shuningdek, ma'lumotlarni minimal yo'qotish bilan saqlash uchun zarur bo'lgan joyni tejaydi.*

PCA usuli ma'lumotlarni ko'p sonli o'lchamlarni ularning teglari bilan 2D yoki 3D fazodagi vizualizatsiya qilish imkonini taqdim etadi [9,10].

Ma`lumotlarni klasterlashtirish

Ma`lumotlarni klasterlash – bu ma'lumotlar to'plamida *kichik guruhlarni aniqlash* yoki *klasterlash* uchun nazoratsiz mashinani o'rganish usuli (4-rasm) [9,10].

Klasterlashning asosiu g'oyasi quyidagicha: Ma'lumotlar to'plamidagi kuzatuvlarni turli xil guruhlarga bo'lish orqali har bir guruh ichidagi kuzatuvlar bir-biriga juda o'xshash va turli guruhlardagi kuzatuvlar bir-biridan mutlaqo farq qilishi lozim. Kuzatishlar o'rtasidagi o'xshashlikni aniqlash uchun mos mezonlarni tanlash kerak. Bugungi kunda klaster tahlili uchun ko'plab usullar ishlab chiqilgan.



4-rasm. *Asosiy komponentlar tahlili jarayoni*

Satrlar va ustunlar bo'yicha klasterlash

Kuzatishlar orasida kichik guruhlarni aniqlash uchun kuzatishlarni xususiyatlar asosida klasterlashimiz mumkin. Shuningdek, biz ustunlar orasidagi kichik guruhlarni aniqlash uchun kuzatishlar asosida xususiyatlarni klasterlashimiz mumkin (xususiyatlarni birgalikda guruhlash). Klasterlash asosan bozor segmentatsiyasi uchun marketing va kompyuterni ko'rish (computer vision)ning ba'zi boshqa sohalarida qo'llaniladi.

Tematik modellashtirish, klasterlash va PCA

Yuqoridagi uchta algoritm, *tematik modellashtirish, klasterlash* va *PCA* nazoratsiz mashinali o`qitish usuli bo`lib, ma'lumotlar to'plamini oz sonli xulosalar bilan soddalashtirish uchun ishlatiladi. Ushbu usullarning asosiy farqi ulardan qanday foydalanishda:

- *LDA (tematik modellashtirish) va PCA usullari o'lchamlarni kamaytirish uchun ishlatilishi mumkin, ammo LDA matn nuqtai nazaridan yaxshiroq aniqlik va tushuntirishni ta'minlaydi.*
- *PCA usuli dispersiyaning bir qismini tushuntiradigan kuzatuvlarning past o'lchamli tasvirini topishga intiladi. Chiqish paramerlari bir-biriga bog'liq emasligi sababli, ularni*

© Elov B.B., Alayev R.H., Aloyev N.R.

klasterlash uchun kirish parametrlari sifatida foydalanish mumkin.

- *Klasterlash klasterlar orasidagi masofani maksimal darajada oshirish orqali kuzatishlar orasidagi kichik guruhlarni topishga asoslangan (klasterlar oldindan ma'lum emas)*

Manfiy bo'lmagan matritsalarni faktorizatsiya qilish (Non-Negative Matrix Factorization, NMF) usuli

NMF – bu matritsani ikkita matritsaga ajratish usuli bo'lib, uchta matritsada ham manfiy elementlar mavjud emas [8,11]. NMF usuli asosan *tavsiya tizimlari, signallarni qayta ishlash* va *bioinformatika* sohalarida qo'llaniladi. NMF usulidagi asosiy g'oya dastlabki kirish matritsasi ikki (odatda) matritsalar (masalan, foydalanuvchilar va filmlar) o'rtasidagi o'zaro ta'sirlardan iborat bo'lgan yashirin xususiyatlar to'plamidan iborat bo'lib, assotsiatsiyalar kirish matritsasidagi og'irliklardir. Manfiy bo'lmagan matritsa faktorizatsiya usulining afzalligi shundaki, biz qanday o'zaro ta'sirlar sodir bo'lishi mumkinligini aniqlashimiz yoki ularni tartiblashimiz.

Yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA) usuli

NLPda LDA usuli yordamida tematik modellashtirish hujjatlardagi so'zlar asosida mumkin bo'lgan mavzularni aniqlash orqali matnli hujjatlar to'plamida **yashirin (latent)** mavzularni aniqlashga imkon beradi [12,13]. Yashirin Dirixle taqsimoti usulida har bir matn hujjat va korpusdagi *har bir so'z, har bir mavzu* va *so'zlar* o'rtasidagi munosabatlar yashirin o'zgaruvchilar yordamida modellashtiriladi.

Korpusdagi har bir hujjat yashirin o'zgaruvchilar (mavzular) bo'yicha **Dirixle taqsimoti** yordamida taqdim etiladi va har bir mavzu barcha hujjatlardagi barcha so'zlar bo'yicha boshqa Dirixle taqsimoti orqali hisoblanadi. LDA usulining umumiy modeli Dirixle taqsimotlarida o'qitilgan generativ ehtimollik modeliga ega **Bayes tizimiga** asoslangan. Bayes taxmini bilan so'zlarga shartli ehtimolliklarni qo'llash orqali kuzatilgan ma'lumotlar (so'zlar) asosida mavzularni modellash mumkin.

LDA usulining quyidagi afzalliklarga ega [13]:

- *Yashirin Dirichlet Allocation (LDA) - boshqa barcha turdagi ilovalar orasida mavzuni modellashtirish uchun eng mashhur yondashuv.*
- *U yaxshi qo'llab-quvvatlanadi va Python, R, Java, C kabi turli xil ramkalarda amalga oshiriladi va ularni joylashtirish juda oson.*

LDA usuli yordamida tematik modellashtirishni amalga oshirish

NLPda tematik modellashtirishni amalga oshirish uchun LDA usulida foydalanilganda, biz kirish matritsasi sifatida so'zlar to'plamini olamiz, chunki bu ehtimollik modeli hisoblanadi. Keyingi qadamda LDA algoritmi matritsani ikkita kichik matritsaga ajratadi:

- *hujjatdan mavzu matrisasi;*
- *so'zdan mavzu matrisasi.*

Ushbu ikkita matrisa shunday optimallashtirilganki, ular bir-biriga ko'paytirilganda so'z matritsasini eng kam **xatolik (lowest error)** bilan hosil qilinadi. Har bir mavzu so'z birikmasi bo'lganligi sababli, har bir kalit so'z har bir mavzuga bir oz vazn beradi. Python tilidagi scikit-learn paketidagi mashhur NLP ilovalardan biri bilan LDA usulini tadbiq qilamiz.

```
# Zarur paketlar va tegishli funktsiyalarni yuklash

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.decomposition import LatentDirichletAllocation

from nlplib import UzTokenizer, UzLemmatizer, stop_words as stopwords

stop_words=stopwords()

# Tajriba uchun mavzular va xususiyatlar sonini o`rnatish

no_features = 1000

no_topics = 10 # Biz 10-namunani tanlaymiz
```

© Elov B.B., Alayev R.H., Aloyev N.R.

```
# LDA usuli uchun faqat qayta ishlanmagan terminlardan foydalanishi mumkin

lemmatizer=UzLemmatizer()

def uz_word_lemmatize(texts):

return lemmatizer.text_lemmatize(texts,False)

# LDA can only use raw term counts for LDA because it is a probabilistic graphical model

tf_vectorizer = CountVectorizer(

max_df=0.95,

min_df=2,

tokenizer=uz_word_lemmatize,

max_features=no_features,

stop_words=stop_words)

tf = tf_vectorizer.fit_transform(documents)

tf_feature_names = tf_vectorizer.get_feature_names_out()
```

Topic 0: son, yil, qaror, Bandi, so‘z
Topic 1: davlat, mablag‘, amalga oshmoq, vazir, yil
Topic 2: qonun, hukumat, yil, bankrotlik, korxona
Topic 3: xizmat, mukofotlamoq, katta, hissa, do‘stlik
Topic 4: yil, qaror, vazir, son, davlat
Topic 5: yil, qaror, son, qaror qilmoq, davlat
Topic 6: osmon, qaror, davlat, transport, viloyat
Topic 7: vazir, maqsad, yil, qaror qilmoq, Oliy
Topic 8: ehtiyoj, samarali, mehnat, qoida, birlashma
Topic 9: qonun, Kir, vazirlik, Oliy, amal

Yashirin semantik taqsimlash (Latent Semantic Allocation,LSA)

Yashirin semantik taqsimlash – NLPda tematik modellashtirish, shuningdek, taqsimlash semantikasi tushunchasidan foydalanadigan o'lchamlarni qisqartirish usuli bo`lib, ushbu usulda ma'nosi o'xshash so'zlar birgalikda uchraydi [14,15].

LSA usuli vositasida hujjat ma'lumotlaridagi yashirin tushunchalarni har bir hujjatdagi so'zlarni *termin hujjat matritsasi (term document matrix, matnning vektor tasviri*) sifatida ifodalanadi. Shuningdek, LSA usuli vositasida IDF matritsasini shakllantirish va keyingi qadamda uni *singular qiymat dekompozitsiyasi (Singular Value Decomposition)dan* foydalangan holda alohida *hujjat-mavzu matritsasi (document-topic matrix)* va *mavzu-termin matritsasi (topic-term matrix)ga* ajratiladi [14,16,17]. Shunday qilib, LSA usulida hujjatlar va terminar o'zlarining past darajali vektor ko'rinishlariga ajratiladi. Shuning uchun *hujjat-hujjat (document-document), hujjat-termin (document-term), termin-termin (term-term)ning* o'xshashligi yoki semantik munosabatlari yashirin semantik o'lchovlar yordamida hisoblanishi mumkin.

Tematik modellashtirishning boshqa NLP algoritmlaridan asosiy farqi shundaki, yashirin semantik fazoda so'rov va hujjat hech qanday umumiy terminlarga ega bo`lmasa ham, **yuqori kosinus o'xshashligiga ega bo'lishi** mumkin.

Ehtimollik yashirin semantik tahlil (Probabilistic Latent Semantic Analysis,PLSA) usuli

LSA usuli bilan solishtirganda, ehtimollik yashirin semantik tahlil usuli (LSA) parchalanish (dekompozisita) uchun SVD o'rniga ehtimollik usulidan foydalanadi [8,18,19]. PLSA usulining asosiy go`yasi hujjat-termin matritsasida biz kuzatadigan ma'lumotlarni yaratishi mumkin bo'lgan yashirin mavzular bilan ehtimollik modelini aniqlashdan iborat.

Xususan, **P(D,W)** modelida, har qanday **d** hujjat va **w** so'zi uchun **P(d,w)** hujjat-termin matritsasidagi yozuvga mos keladi. Har bir hujjat bir nechta mavzularning kombinatsiyasi va har bir mavzu so'zlar to'plami bo'lgan mavzu modellari bilan solishtirganda, LSA ushbu taxmin(bashorat)larga ehtimollik elementini qo'shadi:

- **d** *hujjati berilgan bo'lsa,* **z** *mavzusi ushbu hujjatda* **P(z|d)** *ehtimoli bilan mavjud.*

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

- **z** *mavzusi berilgan bo'lsa,* **w** *so'zi* **z** *dan* **P(w|z)** *ehtimolligi bilan olingan.*

lda2vec chuqur o'rganish modeli

lda2vec –NLPda tematik modellashtirish algoritmi bo'lib, word2vec modelining **skip-gram** arxitekturasini Dirixle optimallashtirilgan siyrak mavzu terminlari bilan aralashtirish orqali *so'zlar va hujjatlarga mos tasvirlar yaratadi* [20.]. Lda2vec modeli asosida kontekstli so'zlarni bashorat qilish uchun vektor so'zini to'g'ridan-to'g'ri ishlatish o'rniga, biz bashorat qilish uchun **kontekst vektoridan** foydalanamiz. Kontekst vektori so'z vektorlari va hujjat vektorlarining yig'indisi sifatida yaratiladi.

Lda2vec modelida so'zlar vektori skip-gram arxitekturasi orqali shakllantiriladi. Hujjat vektori esa hujjatdagi har bir mavzuning og'irliklarini ifodalovchi **hujjat og'irligi vektori** va har bir mavzuni ifodalovchi **mavzu matritsasi** va unga mos keladigan **vektorni joylashtirish** orqali hosil qilinadi. Lda2vec modeli gibrid va kuchli vosita bo`lib, so'zlar uchun so'zlarni joylashtirishni (va kontekst vektor qo'yishlarini) birgalikda o'rganadi va bir vaqtning o'zida mavzu ko'rinishlarini orqali oson va tushunarli o'rganadi.

tBERT usuli

Bert – mavzularni BERT kabi oldindan tayyorlangan kontekstual tasvirlar bilan birlashtirgan NLPda tematik modellashtirish uchun yana bir mavzuni modellashtirish modelidir [21,22]. LDA kabi tipik tematik modellashtirish usuli va Bert Base modellarini birlashtirish tabiiy tilidagi ma'lumotlar to'plamlariga neyron tarmoqlari bo'yicha samarali ishlashni yaxshilaydi. Shuningdek, BERTga mavzular qo'shilishi, ayniqsa, domenga oid holatlarni juda yaxshi hal qilishda yordam berishi ko'rsatilgan.

Tematik modellashtirish va mavzu tasnifi

Tematik modellashtirish va mavzu tasnifi NLP vazifalarining asosiy farqi shundaki, tematik modellashtirish nazoratsiz mashinali o`qitish usuli bo`lsa, mavzu tasnifi esa nazorat qilinadi va qo'lda tayyorlangan ma'lumotlarni talab qiladi:

- *Matnlarni tahlil qilish uchun kam vaqt va soddaroq tahlilga ehtiyoj bo'lmaganda hamda matnlar haqidagi bir nechta mavzu kerak bo'lganda* ***tematik modellashtirish*** *afzaldir.*
- *Boshqa tomondan, matnlar to'plami uchun avtomatik ravishda belgilangan mavzular ro'yxati mavjud bo'lganda va ularni har bir hujjat matnini o'qimasdan avtomatik va aniq belgilash zarurati mavjud bo'lganda* ***mavzularni tasniflash*** *afzalroqdir.*

Tematik modellashtirish ilovalari

NLPda tematik modellashtirishdan tibbiyot, iqtisod va ilmiy tadqiqot sohalarida keng qo'llaniladi. Korpus matnlari tematik modellar yordamida o'qitilgandan so'ng, *his-tuyg'ularni tahlil qilish, hisobotlarni tahlil qilish* va *matnni umumlashtirish* NLP vazifalarida keyingi tahlil qilish uchun kirish parametrlari sifatida foydalanish orqali odamlarga hujjatlar to'plamini tushunish uchun asos yaratadi. Tematik modellar, shuningdek, *tavsiya qiluvchi tizimlar, chatbotlar, qidiruv va virtual yordamchilar* kabi yangi NLP ilovalarida, *qidiruv tizimlarini so'rovlarni kengaytirish, mijozlarga xizmat ko'rsatish va fikr-mulohazalarni ko'rib chiqish* kabi NLP vazidalarida qo'llaniladi.

Xulosa

NLPda tematik modellashtirish – bu katta hajmdagi matnlarni avtomatik ravishda umumlashtirish uchun ishlatilishi mumkin bo'lgan algoritmlar to'plami. Til korpusi matnlarini tahlil qilishda *o'lchov, xususiyatlar* soni juda katta bo'lgan modellarni o'qitishni qiyinlashtiradi va modellarning samaradorligini pasaytiradi. 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.

Yashirin Dirixle taqsimoti NLPda tematik modellashtirish uchun muhim dekompozitsiya usuli bo'lib, hujjatlardan yashirin mavzularni avtomatik ravishda aniqlashda qo`llaniladi. Asosiy komponentlar tahlili shovqinni signaldan ajratish uchun ma'lumotlarni vizual ravishda tushunish uchun o'lchovni kamaytirish usulidir. Klaster – bu xususiyatlar yoki kuzatishlar bo'yicha guruhlashni aniqlashning yana bir nazoratsiz usuli.

© Elov B.B., Alayev R.H., Aloyev N.R.

Tematik modellashtirish va o'lchamlarni qisqartirish quyida keltirilgan algoritmlar bilan amalga oshirilishi mumkin:

- *Manfiy bo'lmagan matritsa faktorizatsiyasi (Non-negative Matrix Factorization);*
- *Yashirin semantik taqsimot (Latent Semantic Allocation, LSA);*
- *Ehtimoliy yashirin semantik tahlil (Probabilistic Latent Semantic Analysis,PLSA);*
- *lda2vec chuqur o'rganish modeli;*
- *tBERT.*

Tematik modellashtirish – bu turli xil sohalarda foydalanish holatlarida qo'llaniladigan ko'p qirrali algoritm. 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. Shuningdek, tematik modellashtirish usullari ijtimoiy tarmoqlardagi foydalanuvchilarning his-tuyg'ularini tahlil qilish imkonini beradi.

Foydalanilgan adabiyotlar

1. 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
2. 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
3. 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
4. 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
5. 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*.
6. Waal, A. de, & Barnard, E. (2008). Evaluating topic models with stability. *Annual Symposium of the Pattern Recognition Association of South Africa*.
7. Hou, C. K. J., & Behdinan, K. (2022). Dimensionality Reduction in Surrogate Modeling: A Review of Combined Methods. In *Data Science and Engineering* (Vol. 7, Issue 4). https://doi.org/10.1007/s41019-022-00193-5
8. 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.
9. 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
10. Qiu, J., Wang, H., Lu, J., Zhang, B., & Du, K.-L. (2012). Neural Network Implementations for PCA and Its Exftensions. *ISRN Artificial Intelligence*, *2012*. https://doi.org/10.5402/2012/847305
11. Wang, J., & Zhang, X. L. (2023). Deep NMF topic modeling. *Neurocomputing*, *515*. https://doi.org/10.1016/j.neucom.2022.10.002
12. 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
13. 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*

© *Elov B.B., Alayev R.H., Aloyev N.R.*

Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009.

14. 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
15. Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1–2).
16. https://doi.org/10.1023/A:1007617005950
17. 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
18. 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
19. 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
20. Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, *42*(1–2).
21. https://doi.org/10.1023/A:1007617005950
22. 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
23. 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
24. 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*.

© *Elov B.B., Alayev R.H., Aloyev N.R.*