

UZBEKISTAN

LANGUAGE & CULTURE

O'ZBEKISTON

TIL VA MADANIYAT

**KOMPYUTER
LINGVISTIKASI**

ISSN 2181-922X

2023 Vol. 2 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 2 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti

Bosh muharrir:

Botir Elov

Bosh muharrir o'rinbosari:

Shahlo Hamroyeva

Mas'ul kotib:

Oqila Abdullayeva

Tahrir kengashi

Shuhrat Sirojiddinov (O'zbekiston), Eshref Adali (Turkiya), Viktor Zaxarov (Rossiya), Vladimir Benko (Slovakiya), Ayrat Gatiatullin (Tataristan), Rinat Gilmullin (Tataristan), Murat O'rxun (Turkiya), Baxtiyor Mengliyev (O'zbekiston), Suyun Karimov (O'zbekiston), Abduvali Qarshiyev (O'zbekiston), O'tkir Hamdamov (O'zbekiston), Tal'at Zuparov (O'zbekiston), Bahodir Mo'minov (O'zbekiston), Faxriddin Nurullayev (O'zbekiston), Zulxumor Xolmanova (O'zbekiston), Dilnoza Muhammadiyeva (O'zbekiston), Muqaddas Abdurahmonova (O'zbekiston), Habibulla Madatov (O'zbekiston), Ilhom Bakiyev (O'zbekiston), Azizaxon Raxmanova (O'zbekiston), Dilrabo Elova (O'zbekiston), Ruhillo Alayev (O'zbekiston), Rasuljon Atamuratov (O'zbekiston), Xolisa Axmedova (O'zbekiston), Zilola Xusainova (O'zbekiston).

Jurnal haqida ma'lumot

“O'zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro'yxatidagi “O'zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o'qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o'zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimai, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to'rt marta chop etiladi.

O'zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e'lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O'zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti. O'zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko'chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz

Website: compling.tsuull.uz

Alisher Navoi'i Tashkent State University of the Uzbek Language and Literature

Chief editor:

Botir Elov

Deputy editor-in-chief:

Shahlo Hamroyeva

Responsible secretary:

Oqila Abdullayeva

Editorial board

Shukhrat Sirojiddinov (Uzbekiston), Eshref Adali (Turkiye), Viktor Zakharov (Russia), Vladimir Benko (Slovakia), Ayrat Gatiatullin (Tataristan), Rinat Gilmullin (Tataristan), Murat Orhun (Turkey), Bakhtiyor Mengliyev (Uzbekistan), Suyun Karimov (Uzbekistan Uzbekistan), Abduvali Karshiyev (Uzbekistan), O'tkir Hamdamov (Uzbekistan), Tal'at Zuparov (Uzbekistan), Bahadir Mo'minov (Uzbekistan), Fakhridin Nurullayev (Uzbekistan), Zulkhumor Kholmanova (Uzbekistan), Dilnoza Muhammadiyeva (Uzbekistan), Muqaddas Abdurakhmonova (Uzbekistan), Habibulla Madatov (Uzbekistan), Ilhom Bakiyev (Uzbekistan), Azizakhan Raxmanova (Uzbekiston), Dirlabo Elova (Uzbekistan), Ruhillo Alayev (Uzbekistan), Rasuljon Atamuratov (Uzbekistan), Kholisa Akhmedova (Uzbekistan), Zilola Khusainova (Uzbekistan).

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz

Website: kompling.tsuull.uz

MUNDARIJA

Shahlo Hamroyeva, Noila Matyakubova

Mashina tarjimasida matnni moslashtirish usullari.....6

Zilola Xusainova

O'zbek tili milliy korpusi qidiruv tizimini optimallashtirishda
lemmatizatsiyadan foydalanish.....20

Shahlo Abdisalomova

O'zbek tilida pronominal anaforani hobbs yondashuvi
asosida hal etish modeli.....38

Botir Elov, Narzullo Alayev, Aziz Yuldashev

Svd va nmf metodlari orqali tematik modellashtirish.....55

Botir Elov, Madina Samatboyeva

Ner: o'zbek tilidagi matnlarda toponim(lar)ni
avtomatik aniqlash modellari.....67

Dilraxon Rustamova

Lingvistik atamalarining so'zligini shakllantirish hamda
terminlarni standartlashtirish asoslari.....85

CONTENT

Shahlo Hamroyeva, Noila Matyakubova

Methods of text alignment in machine translation.....18

Zilola Khusainova

The use of lemmatization in optimizing the search engine
of the national corpus of the uzbek language.....35

Shahlo Abdisalomova

A model for resolving pronominal anaphora in uzbek
based on the hobbs approach.....53

Botir Elov, Narzullo Alayev, Aziz Yuldashev

Thematic modeling using svd and nmf methods.....65

Botir Elov, Madina Samatboyeva

Ner: models for automatic detection of toponym(s)
in uzbek language texts.....84

Dilraboixon Rustamova

Formulation of the dictionary of linguistic terms and
the basis of the standardization of terms.....92

SVD VA NMF METODLARI ORQALI TEMATIK MODELLASHTIRISH

Botir Elov¹,
Narzullo Alayev²,
Aziz Yuldashev³

Annotatsiya.

Tabiiy tilni qayta ishlash (NLP) sohasida tematik modellash-tirish nazoratsiz o'rganish vazifasi bo'lib, uning maqsadi hujjatlar to'plamidan mavhum hisoblangan mavzularni aniqlashdan iborat. Tematik modellash-tirishda ko'p hujjatlarning matn korpusini hisobga olgan holda, matn haqidagi mavhum mavzular aniqlanadi. Tematik modellash-tirish – Machine Learning (ML) uchun nazorat qilib bo'lmaydigan vazifa hisobla-nadi. Ushbu maqolada til korpusi matnlarini SVD va NMF metodlari orqali tematik modellash-tirish masalasi ko'rib chiqiladi.

Kalit so'zlar: til korpusi, tematik modellash-tirish, hujjat-mavzu, atama-mavzu.

Kirish

Bir nechta hujjatlardan iborat katta hajmdagi til korpusi beril-gan bo'lsin. Har bir hujjatni o'qib chiqmasdan (tahlil qilmasdan) tu-rib, berilgan hujjatlar to'plamidagi asosiy mavzularni aniqlash lozim bo'lsin. **Tematik modellash-tirish** orqali til korpusidagi ma'lumot-larni ma'lum miqdordagi *mavzularga ajratiladi* [Vayanskiy, 2020; Churchil, 2022]. **Mavzular** – bu kontekstga o'xshash va hujjatlar to'plamidagi ma'lumotlarni ifodalaydigan *so'zlar guruhi* hisoblana-

¹Elov Botir Boltayevich – texnika fanlari falsafa doktori, dotsent. Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti.

E-pochta: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²Alayev Narzullo Raxmatilloevich – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch doktoranti.

E-pochta: vip.alayev@gmail.com

ORCID: 0009-0009-4625-5539

³Yuldashev Aziz Uyg'un o'g'li – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti o'qituvchisi.

E-pochta: yuldashevaziz@navoiy-uni.uz

ORCID: 0000-0001-6704-2879

di. **M** ta hujjat va **N** ta termin (atama)dan iborat til korpusi uchun “**hujjat - atama**” matritsasi (Document-Term Matrix, DTM)ning umumiy tuzilishi quyida ko’rsatilgan [Pradana, 2020]:

		Terminlar				
		T1	T2	T3	...	TN
Hujjatlar	D1	w11	w12	w13	...	w1N
	D2	w21	w22	w23	...	w2N
	D3	w31	w32	w33	...	w3N
	...	...	...	...	...	...
	DM	wM1	wM2	wM3	...	wMN

1-rasm. M ta hujjat va N ta termin (atama)dan iborat jadval (matritsa)

Berilgan matritsani tahlil qilamiz:

- $D_1, D_2, \dots, D_M$ – M hujjatlar;
- $T_1, T_2, \dots, T_N$ – N atamalar.

“Hujjat-atama” matritsasini to’ldirish uchun keng qo’llaniladigan **TF-IDF** usulidan foydalanamiz.

TF-IDF baholash formulasi

TF-IDF baholash quyidagi tenglama orqali aniqlanadi [Elov, Xusainova, Xudoyberganov, 2022; Elov, Hamroyeva, Xusainova, 2023]:

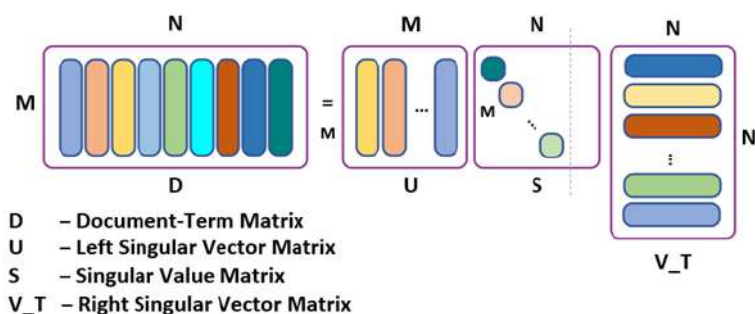
Bu yerda,

- TF_{ij} – D_i hujjatda T_i atamasining uchrash soni;
- df_j – T_j atamasini o’z ichiga olgan hujjatlar soni.

Muayyan hujjatda ko’p ishlatilgan, biroq til korpusda kamdan-kam uchraydigan atama yuqori IDF bahosiga ega bo’ladi. Keyingi qadamda matritsalarini faktorizatsiya qilish usullarini ko’rib chiqiladi.

Singular qiymatlarni ajratish (Singular Value Decomposition, SVD) yordamida tematik modellashtirish

Tematik modellashtirishda SVDning ishlatilishi quyidagi 2-rasmda ko’rsatilgan[Ke, Wang, 2022]:



2-rasm. Singular qiymatlarni ajratish (SVD) yordamida tematik modellashtirish

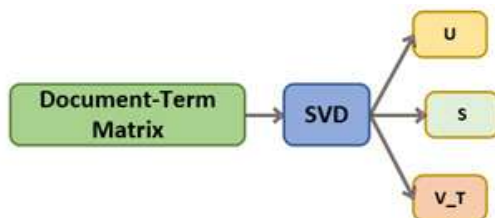
D “hujjat - atama” matritsasidan **SVD** metodi yordamida quyi-dagi 3 ta matritsa hosil qilinadi:

- **U** – *chap singular vektor matritsasi*. Bu matritsa **D.D_T Gram** matritsasining o‘ziga xos bo‘linishi orqali hosil qilinadi. Ko‘p hollarda ushbu matritsa **hujjatlarning o‘xshashlik matritsasi** deb ham nomlanadi. O‘xshashlik matritsasining **i, j**-chi yozuvi **i** hujjat **j** hujjatiga qanchalik o‘xshashligini anglatadi.

- **S** – *Singular qiymat matritsasi*. Ushbu matritsa mavzularning nisbiy ahamiyatini ifodalaydi.

- **V_T** – *o‘ng singulyar vektor matritsasi*. Shuningdek, ushbu matritsa **mavzu matritsasi** deb ham ataladi. Matndagi mavzular ushbu matritsaning satrlari bo‘ylab joylashtiriladi.

SVD metodi orqali **D** “hujjat - atama” matritsasidan 3 ta matritsa (**U**, **S** va **V_T**) hosil qilinadi. Natijada hosil qilingan **V_T** matritsasi qatorlarida mavzular joylashtiriladi.

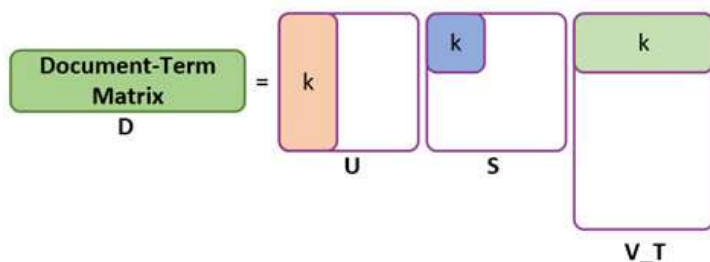


3-rasm. SVD metodi vositasida mavzuni modellashtiruvchi SVD metodi ba’zi hollarda navbatida **Latent Semantic Indexing (LSI)** deb ham nomlanadi.

Qisqartirilgan SVD yoki k-SVD

Aytaylik, **150** ta hujjatdan iborat til korpusi mavjud bo‘lsin. Til korpusini tavsiflovchi **150** xil hujjatni yoki korpus mazmunini ifodalaydigan **10** ta mavzuni o‘qish masalasini ko‘rib chiqamiz. Matn mazmunini yaxshi yetkazib bera oladigan oz sonli mavzularni belgilab olish ko‘pincha foydali hisoblanadi. **k-SVD** metodi vositasida

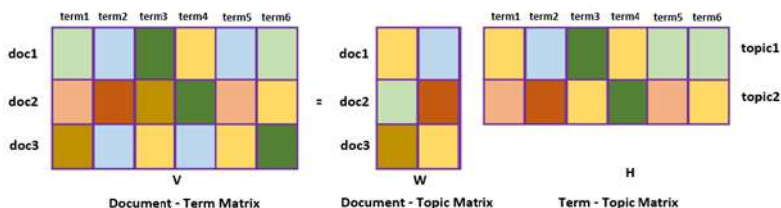
ushbu vazifani bajarish mumkin. Katta o'lchamli matritsalarini o'zaro ko'paytirish katta sondagi murakkab amallarni talab qilganligi sababli, eng katta **k singular qiymatlarni** va ularga mos keladigan **mavzularni** tanlash afzaldir. **k-SVD** metodining ishlashi quyidagi 4-rasmda ko'rsatilgan [Ke, Wang, 2022; Alghamdi, Alfalqi, 2015; Qiang, Qian, Yuan, Wu, 2022]:



4-rasm. k-SVD - eng yaxshi k-darajali approximatsiya

Negativ bo'lmagan matritsali faktorizatsiya (Non-Negative Matrix Factorization, NMF) yordamida tematik modellash

NMF metodining ishlash prinsipi quyidagi 5-rasmda ko'rsatilgan [Kherwa, Bansal, 2020; Lemay, Baek, Doleck, 2021; Wang, Zhang, 2023]:



5-rasm. Negativ bo'lmagan matritsali faktorizatsiya

Bu yerda,

– **W** – **hujjat-mavzu (document-topic matrix)** matritsasi. Ushbu matritsa mavzularning korpusi hujjatlari bo'yicha taqsimotini ifodalaydi.

– **H** – **atama-mavzu (term-topic matrix)** matritsasi. Ushbu matritsa mavzular bo'yicha terminlarning qiymatini ifodalaydi.

NMF metodidagi W va H matritsalarining barcha elementlari manfiy emasligi sababli, korpusga qo'llash birmuncha soddaroq. Shu sababli, NMF metodi orqali natijaning aniqligi biroz yuqori.

NMF – *aniq bo'lmagan matritsalarini faktorizatsiya qilish (non-exact matrix factorization technique)* usuli bo'lib, **W** va **H** matritsalar ko'paytmasi orqali boshlang'ich **V** matritsani aniqlab bo'lmaydi.

Birinchi qadamda **W** va **H** matritsalar tasodifiy tarzda shakllantiriladi. NMF algoritmidagi qadamlar iterativ ravishda bajarilishi natijasida ushbu matritsa qiymatlari yangilanadi va *cost function* (*CF*) deb nomlanuvchi funksiya qiymatini minimallashtiradi. *CF* funksiyasi quyida ko'rsatilganidek, **V-W.H** matritsasining *Frobenius* normasini ifodalaydi:

$$\text{minimize } ||V-WH||_F$$

bu yerda,

- **V** – (*Document - Term Matrix*);
- **W** – (*Document - Topic Matrix*);
- **H** – (*Term - Topic Matrix*).

MxN o'lchovli **A** matritsaning Frobenius normasi quyidagi tenglama bilan aniqlanadi:

$$||A||_F = \sqrt{\sum_{i=1}^M \sum_{j=1}^N |a_{ij}|^2}$$

SVD usuli orqali tematik modellashtirish bosqichlari

SVD usuli orqali tematik modellashtirish uchun quyidagi qadamlarni amalga oshirish lozim.

1-qadam. Mavzularni aniqlashda SVD usulidan foydalanish uchun birinchi qadamda **matn korpusini aniqlab olish** lozim. Quyidagi kod katakchasi kompyuter dasturlash bo'yicha matn bo'lagini o'z ichiga oladi.

text=["Computer programming is the process of designing and building an executable computer program to accomplish a specific computing result or to perform a specific task.",

"Programming involves tasks such as: analysis, generating algorithms, profiling algorithms' accuracy and resource consumption, and the implementation of algorithms in a chosen programming language (commonly referred to as coding).",

"The source program is written in one or more languages that are intelligible to programmers, rather than machine code, which is directly executed by the central processing unit.",

"The purpose of programming is to find a sequence of instructions that will automate the performance of a task (which can be as complex as an operating system) on a computer, often for solving a given problem.",

"Proficient programming thus often requires expertise in several different subjects, including knowledge of the application domain, specialized algorithms, and formal logic.",

“Tasks accompanying and related to programming include: testing, debugging, source code maintenance, implementation of build systems, and management of derived artifacts, such as the machine code of computer programs.”

“These might be considered part of the programming process, but often the term software development is used for this larger process with the term programming, implementation, or coding reserved for the actual writing of code.”

“Software engineering combines engineering techniques with software development practices.”

“Reverse engineering is a related process used by designers, analysts and programmers to understand and re-create/re-implement”]

2-qadam. Matn ma’lumotlari uchun **scikit-learn** paketidan **TfidfVectorizer** sinfini import qilish lozim:

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

Matn korpusi uchun TF-IDF ball (qiymat)lari bilan to’ldirilgan **V** matritsaga ega bo’lish uchun **TfidfVectorizer** sinfidan foydalaniladi:

3-qadam. Yuqorida muhokama qilingan Truncated SVD (k-SVD)dan foydalanish uchun **scikit-learn** paketida **TruncatedSVD** sinfini import qilishingiz kerak:

```
from sklearn.decomposition import TruncatedSVD
```

Barcha zarur modullar import qilganidan so’ng, matndagi mavzularni qidirishga o’tish mumkin.

4-qadam. Ushbu bosqichda **TfidfVectorizer** obyektini yaratish lozim:

```
vectorizer = TfidfVectorizer(stop_words='english',smooth_idf=True)
```

kichik harflar, maxsus belgilarni, nomuhim so’zlarini olib tashlash

```
input_matrix = vectorizer.fit_transform(text).todense()
```

1-4 qadamlar natijasida quyidagi amallar bajarildi:

- matnlar to’plami jamlandi;
- zarur modullar import qilindi;
- Document-Term Matrix matritsasi aniqlandi.

5-qadam. 3-qadamda import qilingan **TruncatedSVD** sinfidan foydalanamiz:

```
svd_modeling= TruncatedSVD(n_components=4, algorithm='randomized', n_iter=100, random_state=122)
```

```
svd_modeling.fit(input_matrix)
```

```
components=svd_modeling.components_  
vocab = vectorizer.get_feature_names()  
6-qadam. Korpus hujjatlariga mos mavzularni aniqlash:  
topic_word_list = []  
def get_topics(components):  
    for i, comp in enumerate(components):  
        terms_comp = zip(vocab,comp)  
        sorted_terms = sorted(terms_comp, key= lambda x:x[1], re-  
verse=True)[:7]  
        topic=""  
        for t in sorted_terms:  
            topic= topic + ' ' + t[0]  
            topic_word_list.append(topic)  
        print(topic_word_list)  
    return topic_word_list  
get_topics(components)
```

7-qadam. Aniqlangan mavzularni va ularni mantiqiy to'g'ri shakllantirilganligini tahlil qilish. SVDdan olingan komponentlar-ni **get_topics()** funksiyasiga parameter sifatida uzatib, *mavzular ro'yxatini* va ushbu mavzularning har biridagi *ommabop so'zlarni* aniqlash:

Topic 1:

code programming process software term computer
engineering

Topic 2:

engineering software development combines practices
techniques used

Topic 3:

code machine source central directly executed intelligible

Topic 4:

computer specific task automate complex given instructions

Yuqorida amalga oshirilgan 7 ta qadam natijasida 4 ta hujjat-dagi mavzular aniqlandi. Aniqlangan mavzularni vizualizatsiya qilish uchun **word cloud** usulidan foydalanish mumkin. Ushbu usul orqali aniqlangan mavzular nisbiy ahamiyatiga ko'ra ko'rsatiladi. Har bir hujjatdagi eng muhim so'z eng katta shrift bilan ajratilgan.

```
from wordcloud import WordCloud  
import matplotlib.pyplot as plt  
for i in range(4):
```

```

wc = WordCloud(width=1000, height=600, margin=3, prefer_horizontal=0.7, scale=1, background_color='black', relative_scaling=0).generate(topic_word_list[i])
plt.imshow(wc)
plt.title(f"Topic{i+1}")
plt.axis("off")
plt.show()

```

Scikit-learn paketidagi NMF sinfi metodlari orqali korpus matnlarini **tematik modellashtirish** mumkin:

```

from sklearn.decomposition import NMF
NMF_model = NMF(n_components=4, random_state=1)
W = NMF_model.fit_transform(input_matrix)
H = NMF_model.components_

```

Keyingi qadamda **get_topics()** metodi orqali **H** matritsasidagi mavzular ro'yxatini aniqlash mumkin:

Topic 1:

code machine source central directly executed intelligible

Topic 2:

engineering software process development used term combines

Topic 3:

algorithms programming application different domain expertise formal

Topic 4:

computer specific task programming automate complex given

Berilgan korpus matnlari uchun SVD va NMF usullari o'xshash mavzular ro'yxatini qaytarishini ko'rish mumkin.

SVD va NMF usullari farqlari

Til korpusi matnlarini tematik modellashtirish uchun ushbu ikkita matritsani faktorizatsiya qilish usullari o'rtasidagi farqlarni keltiramiz:

– *SVD matritsalarini real faktorizatsiya qilish usulidir. SVD usuli natijasida olingan matritsalaridan kirish DTMni qayta hosil qilish mumkin;*

– *Agar korpusga k-SVD usuli qo'llangan bo'lsa, bu holda DTM kirishiga eng yaxshi k-darajali yaqinlashuv amalga oshiriladi;*

Xulosa

Ushbu maqolada til korpusi hujjatlarini SVD va NMF metodlari orqali tematik modellashtirish masalasi ko'rib chiqildi va Python tilidagi sklearn paketi vositalari orqali dasturiy ta'minotni ishlab chiqish ketma-ket qadamlar namoyish etildi. Maqolada M ta hujjat va N ta termin (atama)dan iborat til korpusi uchun DTM (Document-Term Matrix) matrisasini shakllantirish uchun TF-IDF usulidan foydalanildi. Shuningdek, DTM matritsalarini faktorizatsiya qilishning singular qiymatlarni ajratish (Singular Value Decomposition, SVD) va Negativ bo'lmagan matritsali faktorizatsiya (Non-Negative Matrix Factorization, NMF) usullari ko'rib chiqilgan va ularning yutuq va kamchiliklari keltirilgan.

Foydalanilgan adabiyotlar

- Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, 94. <https://doi.org/10.1016/j.is.2020.101582>
- Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, 54(10). <https://doi.org/10.1145/3507900>
- Musthofa Galih Pradana. (2020). Penggunaan Fitur Wordcloud dan Document Term Matrix dalam Text Mining. *Jurnal Ilmiah Informatika*, 8(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*
https://www.academia.edu/105829396/OZBEK_TILI_KORPUSI_MATN-LARI_UCHUN_TF_IDF_STATISTIK_KORSATKICHNI_HISOBLASH
- B.Elov, Sh.Khamroeva, Z.Xusainova (2023). The pipeline processing of NLP. *E3S Web of Conferences* 413, 03011, INTERAGROMASH 2023. <https://doi.org/10.1051/e3sconf/202341303011>
- Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*.
<https://doi.org/10.1080/01621459.2022.2123813>
- Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1). <https://doi.org/10.14569/ijacsa.2015.060121>
- 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>

- Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24). <https://doi.org/10.4108/eai.13-7-2018.159623>
- Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, 2. <https://doi.org/10.1016/j.caeai.2021.100016>
- Wang, J., & Zhang, X. L. (2023). Deep NMF topic modeling. *Neurocomputing*, 515. <https://doi.org/10.1016/j.neucom.2022.10.002>

THEMATIC MODELING USING SVD AND NMF METHODS

Botir Elov¹,
Narzullo Alayev²,
Aziz Yuldashev³

Abstract

In the field of natural language processing (NLP), topic modeling is an unsupervised learning task that aims to identify abstract topics from a set of documents. In thematic modeling, abstract topics about the text are identified by considering the text corpus of many documents. Topic modeling is an intractable task for Machine Learning (ML). This article examines the issue of thematic modeling of language corpus texts using SVD and NMF methods.

Keywords: language corpus, thematic modeling, document-topic, term-topic.

References

- Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. *Information Systems*, 94. <https://doi.org/10.1016/j.is.2020.101582>
- Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, 54(10). <https://doi.org/10.1145/3507900>
- Musthofa Galih Pradana. (2020). Penggunaan Fitur Wordcloud dan Document Term Matrix dalam Text Mining. *Jurnal Ilmiah Informatika*, 8(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*
https://www.academia.edu/105829396/OZBEK_TILI_KORPUSI_MATNLARI_UCHUN_TF_IDF_STATISTIK_KORSATKICHNI_HISOBLASH
- B.Elov, Sh.Khamroeva, Z.Xusainova (2023). The pipeline processing of NLP. *E3S Web of Conferences* 413, 03011, INTERAGROMASH

¹Elov Botir Boltayevich – doctor of philosophy of technical sciences (PhD), associate professor. Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: elov@navoiy-uni.uz

ORCID: 0000-0001-5032-6648

²Alayev Narzullo Raxmatilloevich – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: vip.alayev@gmail.com

ORCID: 0009-0009-4625-5539

³Yuldashev Aziz Uyg'un o'g'li – Teacher of Tashkent State University of Uzbek Language and Literature named after Alisher Navoi.

E-mail: yuldashevaziz@navoiy-uni.uz

ORCID: 0000-0001-6704-2879

2023. <https://doi.org/10.1051/e3sconf/202341303011>
- Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*.
<https://doi.org/10.1080/01621459.2022.2123813>
- Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, 6(1). <https://doi.org/10.14569/ijacsa.2015.060121>
- 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>
- Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24). <https://doi.org/10.4108/eai.13-7-2018.159623>
- Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, 2. <https://doi.org/10.1016/j.caeai.2021.100016>
- Wang, J., & Zhang, X. L. (2023). Deep NMF topic modeling. *Neurocomputing*, 515. <https://doi.org/10.1016/j.neucom.2022.10.002>

Jurnal 2017-yil 26-oktyabrda O'zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro'yxatdan o'tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko'chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 05.09.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. "Cambria" garniturasida.
Shartli b.t. 7,51. Nashr b.t. 7,62.

"O'zbekiston: til va madaniyat" jurnali tahririyatida
tayyorlandi va sahifalandi.
"YASHNOBOD NASHR" bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.