EasyChair Preprint

№ 11004

Topic Modeling Using SVD and NMF

Botir Elov, Narzillo Aloyev and Aziz Yuldashev

October 2, 2023

SVD VA NMF METODLARI ORQALI TEMATIK MODELLASHTIRISH

Botir Elov[1],
Aloyev Narzillo[2],
Aziz Yuldashev[3]

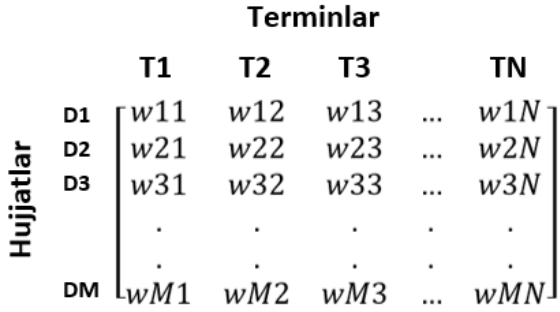
Annotatsiya.

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

Kalit so'zlar: svd, nmf, tematik modellashtirish, hujjat-atama, nlp.

Kirish

Bir nechta hujjatlardan iborat katta hajmdagi til korpusi berilgan bo'lsin. Har bir hujjatni o'qib chiqmasdan (tahlil qilmasdan) turib, berilgan hujjatlar to'plamidagi asosiy mavzularni aniqlash lozim bo`lsin. **Tematik modellashtirish** orqali til korpusidagi ma'lumotlarni ma'lum miqdordagi *mavzularga ajratiladi*[1,2]. **Mavzular** – bu kontekstga o'xshash va hujjatlar to'plamidagi ma'lumotlarni ifodalaydigan *so'zlar guruhi* hisoblanadi. **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[3]:



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

Berilgan matritsani tahlil qilamiz:

---

[1] *Elov 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
[2] *Aloyev Narzillo Raxmatilloyevich* – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch doktoranti.
E-pochta: vip.alayev@gmail.com
ORCID: 0009-0009-4625-5539
[3] *Yuldashev 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

- $\mathbf{D_1, D_2, \ldots, D_M}$ – *M hujjatlar;*
- $\mathbf{T_1, T_2, \ldots, T_N}$ – *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 [4,5]:

$$w_{ij} = TF_{ij} * \log\left(\frac{M}{df_j}\right)$$

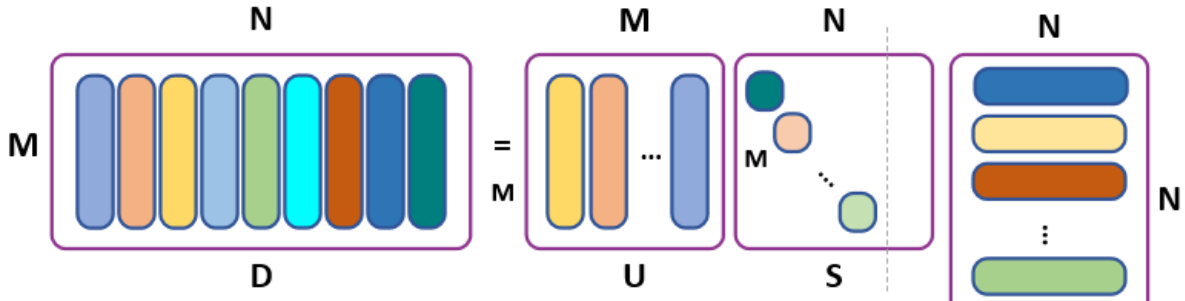
Bu yerda,

- 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 matritsalarni 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[6]:

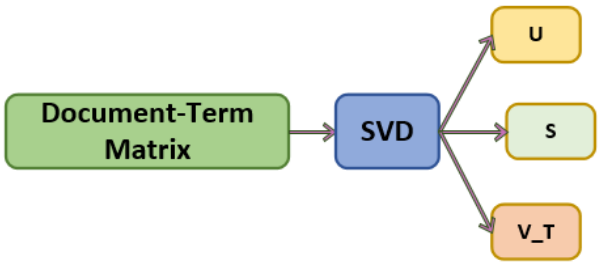


2-rasm. Singular qiymatlarni ajratish (SVD) yordamida tematik modellashtirish

D “hujjat - atama” matritsasidan **SVD** metodi yordamida quyidagi 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.

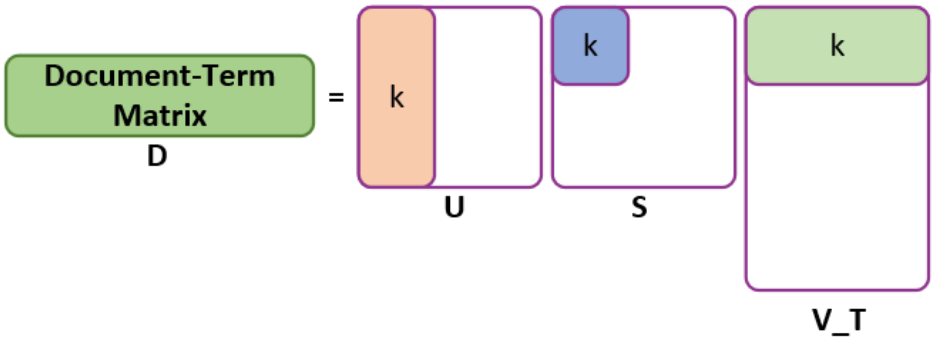


3-rasm. SVD metodi vositasida mavzuni modellashtiruvchidir

SVD metodi baz`i hollarda navbatida **Latent Semantic Indexing (LSI)** deb ham nomlanadi.

Qisqartirilgan SVD yoki k-SVD

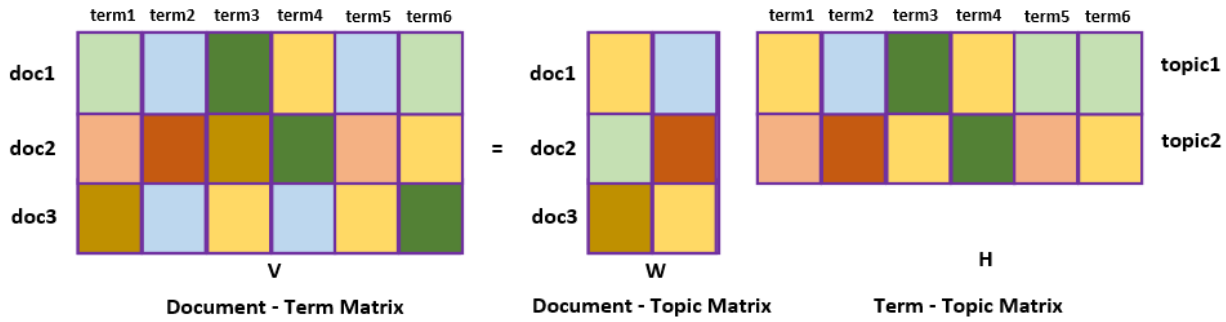
Aytaylik, **150** ta hujjatdan iborat til korpusi mavjud bo`lsin. Til korpusini tavsiflovchi **150** xil hujjatni yoki korpus mazmunini ifodalaydigan **10** ta mavzuni o`qish masalasini ko`rib chiqamiz. Matn mazmunini yaxshi yetqazib bera oladigan oz sonli mavzularni belgilab olish ko'pincha foydali hisoblanadi. **k-SVD** metodi vositasida ushbu vazifani bajarish mumkin. Katta o`lchamli matritsalarni 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[6,7,8]:



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

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

NMF metodining ishlash prinsipi quyidagi 5-rasmda ko’rsatilgan [9,10,11]:



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 matritsalarni 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 normasi*ni ifodalaydi:

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

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:

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

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** paketinida **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 modullarni 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], reverse=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 komponentlarni **get_topics()** funksiyasiga parameter sifatida uzatib, *mavzular ro'yxatini* va ushbu mavzularning har biridagi *ommabop so'zlarni* aniqlanash:

```
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 hujjatdagi 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 matritsalarni real faktorizatsiya qilish usulidir. SVD usuli natijasida olingan matritsalardan 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;*
- *NMF usuli SVD usuliga qaraganda mavzularni aniqlash natijasi yuqori.*

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. Shunigdek, DTM matritsalarini faktorizatsiya qilishniung 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

1. 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
2. Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900
3. Musthofa Galih Pradana. (2020). Penggunaan Fitur Wordcloud dan Document Term Matrix dalam Text Mining. *Jurnal Ilmiah Informatika*, *8*(1).

4. 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*
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. Wang, J., & Zhang, X. L. (2023). Deep NMF topic modeling. *Neurocomputing*, *515*. https://doi.org/10.1016/j.neucom.2022.10.002