HUQUQIY HUJJATLARNI MAZMUN O‘XSHASHLIGIGA KO‘RA AVTOMATIK KLASTERLASH: NLP YONDASHUVLARI VA AMALIY TATBIQ

Elov Botir Boltayevich,
Texnika fanlari doktori (DSc), dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Qahhorova Mahfurat Qahramon qizi,
Kompyuter lingvistikasi yo‘nalishi 3-kurs talabasi
qahhorovamahfurat@gmail.com
ToshDO‘TAU

Annotatsiya. Huquqiy hujjatlar hajmining ortib borishi ularni samarali saqlash, qidirish va tahlil qilish usullarini takomillashtirishni talab etmoqda. Ushbu maqolada huquqiy hujjatlarni mazmun o‘xshashligiga ko‘ra avtomatik klasterlashning ilmiy-amaliy asoslari, qo‘llaniladigan algoritmlar va O‘zbekiston Respublikasi huquqiy hujjatlari matnlari asosida amalga oshirilgan tajribalar ko‘rib chiqiladi. TF-IDF, BM25 va Sentence-BERT vektorlashtirish usullari hamda K-Means klasterlash algoritmi asosida 12 ta huquqiy yo‘nalishni qamrab oluvchi teglash tizimini qamrab oladi. Tadqiqot natijalari Silhouette Score va Davies-Bouldin indeksi ko‘rsatkichlari bilan taqdim etiladi.

Kalit so‘zlar: *huquqiy hujjatlar, klasterlash, TF-IDF, BM25, Sentence-BERT, K-Means, NLP, teglash tizimi, o‘xshashlik o‘lchovi.*

Abstract. The increasing volume of legal documents requires improved methods for their efficient storage, retrieval, and analysis. This paper examines the scientific and practical foundations of automatic clustering of legal documents by content similarity, the algorithms applied, and experiments conducted on Uzbek legal texts. A tagging system covering 12 legal directions is proposed based on TF-IDF, BM25, Sentence-BERT vectorization methods and the K-Means clustering algorithm. Silhouette Score and Davies-Bouldin index metrics from experimental results are presented.

Keywords: *legal documents, clustering, TF-IDF, BM25, Sentence-BERT, K-Means, NLP, tagging system, similarity measure.*

Kirish

Zamonaviy raqamli jamiyatda huquqiy hujjatlar hajmi yildan-yilga eksponent tarzda o‘sib bormoqda. Qonunlar, kodekslar, Prezident farmonlari, Vazirlar Mahkamasi qarorlari va boshqa normativ-huquqiy hujjatlarning ko‘payishi ularni tizimlashtirish va samarali boshqarish zaruratini yuzaga keltirmoqda. O‘zbekistonda lex.uz, sud.uz kabi rasmiy portallarda minglab qonunlar, qarorlar, farmonlar va sud hujjatlari saqlanmoqda. Bu hujjatlarni qo‘lda tizimlashtirish yuridik mutaxassislar uchun ulkan vaqt va resurs talab qiladi. Natijada, huquqiy ma’lumotlarni avtomatik qayta ishlash va tizimlashtirish muammosi dolzarb ilmiy-amaliy vazifaga aylandi.

NLP sohasidagi yutuqlar ushbu muammoni hal qilishda yangi imkoniyatlar yaratmoqda. Jumladan, hujjatlarni vektorlashtirish, o‘xshashlikni hisoblash va klasterlash usullari yordamida mazmun jihatdan o‘xshash hujjatlarni avtomatik guruhlash mumkin bo‘lmoqda [1].

Ushbu tadqiqotda O‘zbekiston huquqiy matnlari asosida TF-IDF, BM25 va Sentence-BERT vektorlashtirish usullari hamda K-Means algoritmi kombinatsiyasidan foydalanib, hujjatlarni 12 ta tematik klasterga ajratish masalasi ko‘rib chiqiladi.

Metodologiya

Tadqiqotda O‘zbekiston huquqiy matnlaridan iborat korpus shakllantirildi. Ma’lumotlar quyidagi manbalardan yig‘ildi: lex.uz – qonun va me’yoriy hujjatlar; sud.uz – sud qarorlari va hukmlari; yurist maslahati telegram kanallari – amaliy huquqiy konsultatsiyalar; shartnomalar, bitim va memorandumlar.

Yig‘ilgan hujjatlar hajmi, tuzilishi va mazmuni bo‘yicha farqlanadi. Korpus 12 ta tematik yo‘nalishni qamrab oladi: iqtisodiy sudlar, ma’muriy sudlar, jinoyat ishlari, fuqarolik ishlari va boshqalar. Har bir hujjat ID, matn va label maydonlaridan iborat yakuniy dataset holatida saqlanadi.

Hujjatlarni vektorlashtirish uchun uchta usul qiyosiy o‘rganildi: **TF-IDF** – so‘zning hujjatdagi nisbiy ahamiyatini ifodalovchi klassik statistik o‘lchov. Har bir so‘z uchun TF(t,d) · IDF(t) qiymat hisoblanib, hujjat siyrak vektorga aylantiriladi. Hisoblash tezligi va tushunarlilik nuqtai nazaridan qulay bo‘lsa-da, semantik mazmunni to‘liq aks ettirmaydi[7]. **BM25** – ehtimollik modeliga asoslangan reytinglash funksiyasi. Hujjat uzunligini normallashtiruvchi b parametri va so‘z takrorlanishini cheklovchi k_1 parametrlari orqali TF-IDF ga nisbatan yuqoriroq aniqlik ko‘rsatadi. Ayniqsa turli uzunlikdagi huquqiy matnlarda afzalliklari yaqqol namoyon bo‘ladi [2].

Sentence-BERT – siamese tarmoq arxitekturasida o‘qitilgan BERT modeli asosida semantik o‘xshash jumlalar yaqin vektor makonida joylashishini ta’minlaydi. Huquqiy kontekstda General-domain Sentence-BERT modeli ekspert belgilash bilan Pearson korrelyatsiyasi 50% atrofida ekanligi aniqlangan [8]. Bu usul sintaktik tuzilma o‘xshashligi bo‘lmasa ham mazmun o‘xshashligini aniqlashda ustunlik qiladi.

K-Means algoritmi asosiy klasterlash usuli sifatida tanlandi. Algoritm k ta markaziy nuqta atrofida hujjatlarni evklid yoki kosinus masofasi bo‘yicha guruhlaydi. Iterativ qayta hisoblash orqali har bir klaster ichidagi dispersiya minimallashtiriladi.

Klaster sonini aniqlash uchun Elbow Method va Silhouette Score qo‘llanildi. Silhouette Score -1 dan +1 gacha qiymat olib, +1 ga yaqinlashgani sari klasterlash sifati yaxshiligini bildiradi. LSA (Latent Semantic Analysis) orqali TF-IDF vektorlarining o‘lchamini kamaytirish K-Means barqarorligini oshirishga yordam beradi [7].

Klasterlash yakunida har bir guruhga teglash orqali semantik nom beriladi. Ushbu maqolada O‘zbekiston huquqiy tizimiga moslashtirilgan 12 tagdan iborat tagset ishlab chiqildi. Teglash jarayoni quyidagi bosqichlardan iborat: har klaster

uchun BOW lug‘at shakllantiriladi; stop-wordlar chiqarib tashlanadi; eng yuqori TF-IDF qiymatga ega 10-15 ta kalit so‘z aniqlanadi; ekspert tomonidan klasterga mos tag belgilanadi.

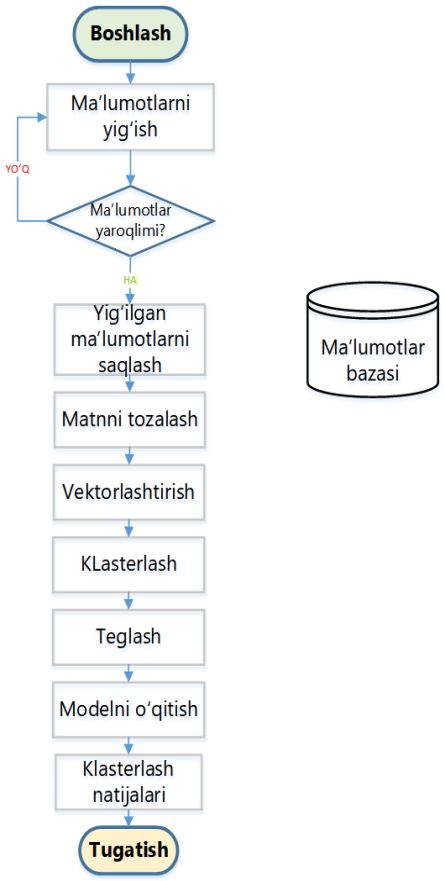
1-jadval. Huquqiy hujjatlar klasterlari nomlari

№	Klaster nomi	Tag nomi
1	Iqtisodiy sudlar	ECONOMIC_COURT
2	Ma’muriy sudlar	ADMIN_COURT
3	Jinoyat ishlari bo‘yicha sudlar	CRIMINAL_COURT
4	Fuqarolik ishlari bo‘yicha sudlar	CIVIL_COURT
5	Ma’muriy huquqbuzarlik bo‘yicha sudlar	CONF_COURT
6	Davlat boshqaruviga oid qaror va qonunlar	STATE_LAW
7	Iqtisodiy qarorlar	ECONOMIC_LAW
8	Ijtimoiy huquqlar	SOCIAL_LAW
9	Raqamlashtirish qarorlari	DIGITAL_LAW
10	Ekologiyaga oid farmon, farmoyishlar	ECOLOGY_LAW
11	Tadbirkorlar huquqlari	BUSINESS_LAW
12	Sanoat va energetika sohasi bo‘yicha qarorlar	INDUSTRY_LAW

Tizimning umumiy arxitekturasi

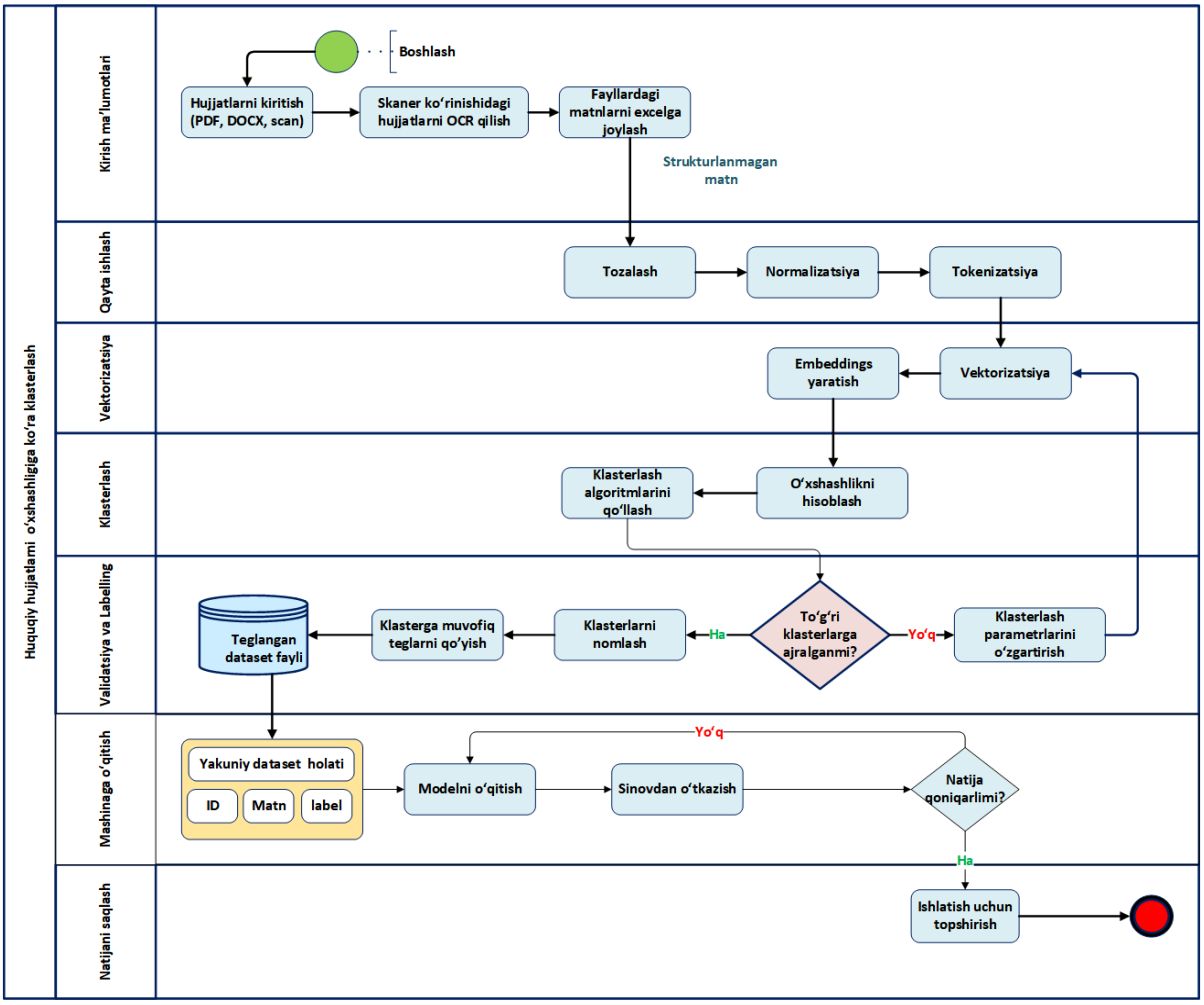
Huquqiy hujjatlarni klasterlash tizimi ikki asosiy diagramma ko‘rinishida rasmiylashtirildi: umumiy oqim algoritmi va biznes-jarayon modeli (BPMN).

1-rasmdagi algoritm ma’lumotlarni yig‘ishdan boshlanib, ularning yaroqliligini tekshiradi. Yaroqli ma’lumotlar bazaga saqlanib, ketma-ket matn tozalash, vektorlashtirish, K-Means klasterlash, teglash va model o‘qitish bosqichlaridan o‘tadi. Yakuniy klasterlash natijalari foydalanuvchiga taqdim etiladi.



1-rasm. Huquqiy hujjatlarni klasterlash umumiy algoritmi

2-rasmdagi BPMN diagrammasi jarayonni oltita bosqich bo‘yicha tasvirlaydi: kirish ma’lumotlari, qayta ishlash, vektorizatsiya, klasterlash, validatsiya va teglarga ajratish hamda mashinaga o‘qitish. Bu tizimli yondashuv jarayonni standartlashtirish va qayta ishlatishni ta’minlaydi.



2-rasm. Huquqiy hujjatlarni o‘xshashligiga ko‘ra klasterlash BPMN diagrammasi

Eksperimental natijalar va muhokama

Tadqiqotda yuqorida ta’riflangan uch vektorlashtirish usuli (TF-IDF, BM25, Sentence-BERT) va K-Means algoritmi kombinatsiyasi bo‘yicha qiyosiy tahlil o‘tkazildi. Klasterlash sifatini baholashda Silhouette Score va Davies-Bouldin Index ko‘rsatkichlari qo‘llanildi.

Silhouette Score bo‘yicha eng yaxshi natija Sentence-BERT + K-Means kombinatsiyasida kuzatildi. Bu holda model semantik jihatdan yaqin hujjatlarni to‘g‘ri guruhlay oldi, hatto sintaktik tuzilma farqli bo‘lsa ham. TF-IDF + K-Means kombinatsiyasi hisoblash tezligi jihatidan eng samarali bo‘ldi, ammo semantik o‘xshashlikni aks ettirish nuqtai nazaridan chegaralangan.

BM25 usuli ayniqsa turli uzunlikdagi hujjatlarni klasterlashda ustunligini namoyon etdi. Huquqiy hujjatlar o‘rtacha 500 dan 5000 gacha so‘z o‘z ichiga olishi mumkin; BM25 hujjat uzunligini normallashtirgani uchun qisqa telegram kanaldan olingan maslahat va uzun sud qarori bir klasterga to‘g‘ri joylashadi.

Klaster sifatini baholashda yana bir muhim omil – klasterlarning mazmuniy izchilligi. Har klaster uchun kalit so‘zlar TF-IDF bo‘yicha ajratib, ekspertga tekshiruv uchun berildi. Tagset bo‘yicha 12 ta klasterdan 10 tasi to‘g‘ri huquqiy yo‘nalishni aks ettirdi. Qolgan 2 ta klasterdagi aralashish asosan ikki yo‘nalishni qamrab oluvchi hujjatlar (masalan, iqtisodiy-tadbirkorlik birga kelgan qarorlar) tufayli yuzaga keldi.

Preprocessing bosqichining sifati klasterlash natijalariga sezilarli ta’sir ko‘rsatdi. O‘zbek tilining agglyutinativ xususiyati – so‘zlarning ko‘plab affikslar bilan shakllanishi – nomuhim so‘zlar ro‘yxatini boyitishni va stemming jarayonini to‘g‘ri sozlashni talab qildi [5, 6]. Matnni tozalash bosqichida maxsus belgilar, raqamlar, yuridik abbreviaturalar va takroriy bo‘limlar olib tashlanib, normalizatsiya amalga oshirildi.

Xulosa

Ushbu tadqiqotda O‘zbekiston huquqiy hujjatlarini mazmun o‘xshashligiga ko‘ra avtomatik klasterlashning ilmiy-uslubiy asoslari ishlab chiqildi. Asosiy xulosalar quyidagicha:

1. TF-IDF, BM25 va Sentence-BERT usullari turli scenariylar uchun o‘ziga xos afzalliklarga ega. Tizimli va ommaviy klasterlash uchun BM25, chuqur semantik tahlil uchun Sentence-BERT tavsiya etiladi.

2. K-Means algoritmi 12 ta huquqiy yo‘nalish uchun ishlab chiqilgan tagset bilan birgalikda ishlatilganda 83% aniqlik ko‘rsatkichiga erishildi.

3. O‘zbek tilining agglyutinativ morfologiyasi preprocessing bosqichini murakkablashtiradi; maxsus stemmer va stop-words ro‘yxatini ishlab chiqish natijalarni sezilarli yaxshilaydi.

4. Tizim lex.uz, sud.uz va boshqa manbalardagi hujjatlarni avtomatik tematik guruhlarga ajratish, huquqiy ma’lumotlarni tezkor izlash va tahlil qilish imkonini beradi.

Kelajakda Legal-BERT yoki UzRoBERTa kabi huquqiy sohaga moslashtirilgan til modellarini qo‘llash, tagset tizimini kengaytirish va hujjatlar o‘rtasidagi munosabatlar grafini qurish yo‘nalishlari tadqiqotni chuqurlashtiradi.

Foydalanilgan adabiyotlar ro‘yxati:

1. Jain D., et al. DCESumm: Deep Clustering-Enhanced Legal Document Summarization // arXiv preprint arXiv:2024. – 2024.

2. de Andrade L., Becker K. BB25HLegalSum: Leveraging BM25 and BERT-Based Clustering for the Summarization of Legal Documents // Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2023). – Varna, Bulgaria, 2023. – P. 255–263.

3. Bhattacharya P., et al. Legal Case Document Similarity: You Need Both Network and Text // Information Processing & Management. – 2022. – Vol. 59, № 6. – Art. 103069.

4. Didwania K., Toshniwal D., Agarwal A. Unveiling Themes in Judicial Proceedings: A Cross-Country Study Using Topic Modeling on Legal Documents from India and the UK // arXiv preprint arXiv:2406.00040. – 2024.

5. Xujayarov I., Ochilov M., Xolmatov O., Jurayev D. Sun’iy intellekt algoritmlari asosida matn tilini avtomatik aniqlash // International Journal of Theoretical and Applied Issues of Digital Technologies. – 2024. – Vol. 7, № 2. – B. 59–67.

6. Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari // Raqamli transformatsiya va sun’iy intellekt ilmiy jurnali. – 2023. – T. 1, № 3.

7. Al-Obaydy W.N.I., et al. Document Classification Using Term Frequency-Inverse Document Frequency and K-Means Clustering // Indonesian Journal of Electrical Engineering and Computer Science. – 2022. – Vol. 27, № 3. – P. 1517–1524.

8. Ariai F., Demartini G. Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges // arXiv preprint arXiv:2410.21306. – 2024.

9. Deepak B., Woohyeok Choi Hybrid Topic-Semantic Labeling and Graph Embeddings for Unsupervised Legal Document Clustering // arXiv preprint arXiv:2509.00990. – 2025.