MATNLI HUJJATLARNI QAYTA ISHLASHNI TASNIFLASH DASTURLARI

Choryorqulov G‘iyos Husan o‘g‘li *- assistent*
O‘zbekiston Milliy universitetining Jizzax filiali, O‘zbekiston
choryorqulovg31@gmail.com
Qorjovova A’lo Yaxshiliq qizi *- talaba*
O‘zbekiston Milliy universitetining Jizzax filiali, O‘zbekiston
aloqorjovova23@gmail.com

Annotatsiya. Ma’lumotlar intellektual tahlili tizimlarining bir yo‘nalishi sifatida qaraladigan Text Mining matn va matnli hujjatlarni tahlillash bilan bog‘liq masalalarni (klassifikasiya, klasterizasiya, asosiasiya kabilar) hal qilishga yo‘naltirilgan mexanizmlari – dasturiy vositalarini tahlil qilish o‘rinli.Ushbu maqolada ushbu daturiy vositalar yoritib berilgan.

Kalit so‘zlar: word2vec, Optik belgilarni aniqlash (OCR), morfologik lug‘at.

Kirish:

Matnli ma’lumotlarni qayta ishlash, matn tahrirlashda juda muhim bo‘lib, matndagi xatolar va noaniq ma’lumotlarni bartaraf qilish yoki matning ko‘rsatilgan ma’noga to‘g‘ri kelishini ta’minlash yordam beradi.

Ma’lumotlar intelektual tahlili murakkab va ko‘p bosqichli jarayon bo‘lib, o‘zida turli bosqichlarda ma’lumotlarni yig‘ish, dastlabki ishlov berish va tahlil qilish jarayonlarini mujassamlashtiradi. Ma’lumotga dastlabki ishlov berish fazasida turli formatdagi ma’lumotlarni tartiblash, bir formatga keltirish uchun bir qator algoritmlarni qo‘llashni talab qiladi.

Matn bilan ishlovchi tizimlar.

Matn bilan ishlash jarayonini to‘g‘ri va samarali tashkil etilishi axborot oqimini boshqarish va va egalik qilish imkonini beradi. Quyida jahon tan olgan bir guruh tizimlar tavsiflangan.

1.word2vec - Google korporasiyasi uchun Tomas Malikov guruhi tomonidan ishlab chiqilgan tadqiqot loyihasi bo‘lib, o‘zini o‘qituvchi "Continuous Bag-of-Words" (CBOW) i "Skip-gram model" (SG) neyron to‘rlari algoritmlari asosida amalga oshirilgan.

2.Ularning yordamida kontekstda so‘zning yozilishi asosida “lug‘at” shakllantiriladi. Mazkur lug‘atning hajmi o‘quv tanlanmasi hajmini belgilab beradi.

Bu esa tasniflash aniqligini oshirishda asosiy omil hisoblanadi. word2vec vositasining bir qator klonlari mavjud, bulardan biri https://github.com/dav/word2vec - loyihasi bo‘lib, mazkur vosita uzliksiz so‘zlar paketi arxitekturasi hamda so‘zlarninng vektor ifodasini hisoblashga yo‘naltirilgan skip-grammlarni amalga oshirishni nazarda tutadi. Vositadan tabiiy til matnini tasniflash va ishlov berishda tadqiqot maqsadlarida foydalanilishi mumkin. Ishlov berilgunga qadar rus orfografiyasini tuzatish tizimi rus orfografisi qoidalariga asoslangan ruscha matnni grammatik tadqiq qilishni amalga oshiradi [11]. Tizim o‘zida quyidagi imkoniyatlarni mujassamlashtiradi: - tokenlar va ularning turlari (so‘z, tinish belgilari, miqdor, teg belgisi va h.k.)larni ajratish; - lug‘at so‘zlarning morfologik tahlilini amalga oshirish; - tanib olinmagan so‘zlar uchun tahminlar qurish; - ko‘pso‘zli birikmalarni tahlil qilish. Apache OpenNLP (The Apache Software Foundation, Incubator) – Apache nomi ostida tabiiy til matniga mashinali ishlov berish bilan bog‘li ochiq kodli vosita hisoblanadi.

3.Mazkur vosita ham o‘zida tokenlash, gaplarni ajratish, nutq qismlarini tahlillash, kelishiklarni aniqlash, matnni ajratish va kesishuvchi so‘zlarni tahlillash kabi imkoniyatlarni mujassamlashtirgan. OpenNLP o‘zida mashinali o‘qitish bilan bog‘liq Java-vositalarining keng qurollarini jamlaydi. Link Grammar Parser (John Lafferty, Daniel Sleator, Davy Temperley, Carnegi Melon University, USA) – ingliz tilining sintaksis tahlillagichi bo‘lib, 60000 so‘z ko‘rinishlariga ega lug‘atdan foydalanadi C tilida amalga oshirilgan va Unix ga mo‘ljallangan. Windows API32 uchun ishlab chiqilgan versiyasi ham mavjud. Konsolli interfeysga ega, kiruvchi ma’lumotlar klaviyatura yordamida ko‘lda yoki ASCII-faylda paketli ishlov berish uchun kiritiladi[3].

4.Cibola/Oleada loyihalari (Computing Research Laboratory (CLR), New-Mexico State University, USA) – Unicode da berilgan matnlarning keng spektrdagi lingvistik tahlilini amalga oshiradi.

5.Tizimning komponentalari multitilli matnlar bilan (MUTT) ishlashga yo‘naltirilgan 16 tilli statik tahlilga ega, avtomatik tarjima mexanizmli modullardan iborat. Tizimning barcha komponentalari SunOs va Solaris uchun X11 Window System muhitlari uchun amalga oshirilgan. Russian Morphological Dictionary (S.Sikorskiy) – rus tilidagi matnlarni sintaksis va morfologik tahlilini amalga oshirish uchun mo‘ljallangan.

6.Morfologik lug‘atga ega, lug‘at hajmi 120 ming so‘zdan iborat. Tizim SWI-Prolog muhitida Windows uchun yaratilgan. Mystem (Ilya Segalovich, Vitaliy Titov,

kompaniya Yandex) – rus tilidagi matnlarni morfologik tahlilini amalga oshiruvchi ihcham, juda tezkor va tekin ilova hisoblanadi. Tizim konsolli rejimda ishlaydi[4].

7.MyStem tizimi lug‘atda mavjud bo‘lmagan so‘zlarni ham gippotetik taqsimotini qura oladi. LingSoft (LingSoft, Finlyandiya) – bir qator tillar guruhi uchun matnlarni grammatik tafsiflash, morfologik tahlillash, normallashtirishni amalga oshiroladigan komponenta hisoblanadi.

8.StarLing (S.A.Starostin) – uzun matnli multi tilli matnlar, transklipsiya belgilari bilan ishlovchi tizim bo‘lib, qulay qidiruv tizimi, so‘z tuzilishlarini tahillash va sintezlashimkoniyatlarini o‘zida mujassamlashtirgan ma’lumotlar bazasini boshqarish tizimi hisoblanadi.

9.Tizim o‘ziga Ojegova (zaliznia.exe) va Zaliznyaka (http://starling.rinet.ru/download/zaliznia.exe) larning DBF-formatdagi lug‘atlarini ham yuklab olish imkoniyatiga ega. on-line rejimda saytda turli tillar uchun etimologik baza mavjud[5].

10.MonoConc/ParaConc (Michael Barlow, Dept of Linguistics, Rice University, Texas, USA) – ko‘p tilli matnlarni tahlil qilishga yo‘naltirilgan tijorat dasturi.

11.Bepul ko‘rinishdagi demo-versiyalari ham mavjud bo‘lib, ularda funksionallik cheklangan. WordSmith Tools (Mike Scott, 2010, School of English, University of Liverpool) – matnlardagi so‘zlarni ishlatilish formasini tadqiq qilishga yo‘naltirilgan ko‘p tilli ilova bo‘lib, olingan so‘z va so‘z yasovchilarning tartiblangan ro‘yxatini shakllantirib beradi. Turli formatdagi matnli hujjatlar bilan ishlash imkoniyatiga ega - PDF, MS Word, HTML, XML yoki SGML.

12.TextAnalyst 2.0 ("MikroSistemы" ilmiy ishlab chiqarish innovasion markazi).

13. O‘zida bir qator komponentalarni mujassamlashtiruvchi tadqiqot natijalari jamlanmasi bo‘lib, uning yordamida rus va ingliz matnlarini avtomatlashtirilgan tahlilini amalga oshiradi. Shu bilan birga o‘zida matnni semantik tahlil qiluvchi komponentasiga ham ega. Matnlarni tahrirlash uchun ilova taklif etilgan bo‘lib ulardan demo-versiya sifatida foydalanish mumkin[6].

14. netXtract (Relevant Software Inc.) hamda Textual Analysis Computing Tools (TACT) (Library Electronic Text Resource Service Indiana University, USA) vositalari o‘zida matnli hujjatlarni tahlillash komponentasini mujassamlashtirib, veb-matnlarga ishlov berishga yo‘naltirilgan. WordTabulator (Logichev S.V., 1997-2016) - MS Windows muhitida matnlarni tahlillash uchun ishlab chiqilgan vosita.

15. Dastur ANSI, UTF-8 yoki HTML formatdagi matnlarni dastlabki ma’lumot sifatida qabul qilishi mumkin. Oracle Text (Oracle) orakl ma’lumotlar bazasida, fayllarda va internetda saqlanadigan matn va hujjatlarni indekslash, qidirish va tahlil qilish uchun SQL-standart dan foydalanadi[7]. Oracle Text hujjatlarda lingvistik tahlilni amalga oshirishi, shuningdek, kalit so‘zlarni qidirish, kontekst so‘rovlari, buxgalteriya operasiyalari, namunalarni topish, aralash tematik so‘rovlar, HTML/XML qismni qidirish va hokazolarni o‘z ichiga olgan turli xil strategiyalardan foydalangan holda matnni qidirishni amalga oshirishi mumkin. U qidiruv natijalarini turli formatlarda, shu jumladan formatlanmagan matnda, HTML belgilashlar bilan va hujjat asl formatida ko‘rsatishi mumkin. Oracle Text bir nechta tillarni qo‘llab-quvvatlaydi va qidiruv sifatini oshirish uchun ilg‘or texnologiyalaridan foydalanadi. Oracle Text shuningdek, klassifikasiya, klasterlash va ma’lumot vizual metaforalarini qo‘llab-quvvatlash kabi ilg‘or xususiyatlarni taklif etadi[8].

Optik belgilarni aniqlash (OCR) dasturiy tizimining ishlashi.

Optik belgilarni aniqlash (OCR) dasturi chop etilgan hujjatlarni kompyuterga avtomatik kiritish uchun mo‘ljallangan. Ko‘pincha foydalanuvchilar amalda Cognitive Technologies Ltd tomonidan ishlab chiqarilgan ABBYY FineReader optik belgilarni aniqlash tizimi va CuneiForm optik belgilarni aniqlash tizimidan foydalanadilar. Ikkala tizim ham taxminan bir xil imkoniyatlarga ega va deyarli har qanday shriftda (ieroglif va arabchadan tashqari) terilgan matnlarni oldindan tayyorgarliksiz tanib olish uchun mo‘ljallangan. Dasturlarning o‘ziga xos xususiyati belgilarni aniqlashning yuqori aniqligi va bosib chiqarish nuqsonlariga nisbatan past sezgirlikdir[9].

OCR dasturiy tizimlari quyidagicha ishlaydi. Aytaylik, sizda matnli qog‘oz hujjati bor murakkab tuzilish, ya'ni. matndan tashqari, hujjatda jadvallar, diagrammalar, rasmlar va boshqalar mavjud. Matnni matn protsessoridan foydalanib tahrirlashingiz kerak. Bunday muammoni hal qilish uchun siz hujjatni skanerga joylashtirishingiz va uning elektron nusxasini yaratishingiz kerak, bu hujjatning grafik tasviridir.

Keyinchalik, rasmni matnga aylantirishingiz kerak. Bu bosqich juda mas'uliyatli, chunki skanerlash natijasi faqat u yoki bu grafik formatdagi faylda saqlanishi yoki matnga emas, balki har qanday grafik muharrirga ishlov berish uchun yuklanishi mumkin bo‘lgan tasvirdir. Tasvirni to‘g‘ridan-to‘g‘ri matn muharririga kiritishingiz mumkin. Biroq, matn ilovalari uchun rasm bo‘linmas element bo‘lib, uni aniqlab bo‘lmaydi. Shunday qilib, rasmda matn mavjud bo‘lsa ham, uni matn

muharriri yordamida tahrirlash mumkin emas. Shuning uchun, birinchi navbatda, belgilar tasvirini matnga shunday aylantirish kerak, ya'ni. matn muharrirlarida qayta ishlash uchun mavjud belgilar ketma-ketligiga.

FineReader va CuneiForm dasturiy ta'minot tizimlari tasvirlarni matnga aylantirish muammolarini hal qilishga imkon beradi va ulardan foydalanish uchun yetarlicha kuchli komplekslar shaklida taqdim etiladi[10].

OCR tizimida matnni aniqlash jarayoni.

OCR tizimlari tomonidan matnni aniqlash jarayoni quyidagicha:

Birinchidan, hujjatning grafik tasvirini olishingiz kerak, uni ikki usulda - hujjatni skanerlash yoki fayldan rasmni yuklash orqali amalga oshirish mumkin. Ilova muhiti tasvirlarni turli masshtablarda ko‘rish, shuningdek, ular ustida ba'zi o‘zgarishlarni amalga oshirish, xususan, ularni aylantirish va aylantirish imkonini beradi.

Hujjatlarni qayta ishlashning keyingi bosqichida ular belgilanadi. Ushbu operatsiyaning maqsadi OCR tizimiga matnning tasvirda qanday joylashishini aytib berishdir.

Avtomatik belgilash natijalari qo‘lda tuzatilishi mumkin - matn bloklarini yaratish va o‘chirish, ularni ko‘chirish, o‘lchamlarini o‘rnatish, qo‘shni bloklarga bo‘lish, bloklarning to‘rtburchaklarini biriktirish yoki kesish orqali bloklarni ko‘pburchak qilish va hokazo. Bloklarga ketma-ket raqamlar berilishi mumkin, agar siz murakkab formatlangan matnni oddiy matnga aylantirishingiz kerak bo‘lsa, bu juda qulaydir[11].

Noaniq yoki bir xil bo‘lmagan fonda chop etilgan matnlar uchun adaptiv skanerlash qo‘llaniladi, bu esa unumdorlikning biroz pasayishi hisobiga harflar konturlarini aniqlashning aniqligini oshirishga imkon beradi.

Kam kontrastli, zaif bosilgan hujjatlarni tanib olishda yorqinlik, kontrast va qora va oq nuqta chegarasi kabi sozlamalarni sozlash orqali tanish sifatini yaxshilash mumkin.

Tartib va tanib olish imkoniyatlari OCR tizimi matnni to‘g‘ri bloklarga bo‘lish va uni tanib olish imkoniyatiga ega bo‘lishi uchun tuzilgan. Xususan, tan olingan matn qaysi tilda (tillarda) yozilganligini ko‘rsatishingiz kerak.

Operativ ish tan olingan matn bilan OCR tizimi o‘zining matn muharririni birlashtirib, o‘rnatilgan Windows dasturi WordPad dasturini eslatadi. Matn muharriri shriftlar va qirralar, ustki va pastki belgilar, jadvallar, ustunlar, matn ustida suzuvchi ramkalar kabi matnni formatlashning asosiy xususiyatlarini qo‘llab-quvvatlashga qodir. Shubhali so‘zlar ma'lum bir fon bilan tanilgan matnda ta'kidlanadi va matn

muharriri shubhali so‘zlarni tezda topish uchun vositalarni taqdim etadi, bu esa tan olingan matnni ko‘rish va tahrirlashni sezilarli darajada osonlashtiradi[12].

OCR tizimlarining ba'zi versiyalari, masalan, ABBYY FineReader Corporate Edition o‘rnatilgan hamkorlik vositalarini o‘z ichiga oladi. Tarmoqda ishlash imkoniyatini amalga oshirish uchun har bir kompyuterda dasturning alohida nusxasi o‘rnatilishi kerak.

Bunday holda, ishni bir nechta kompyuterlarda bir xil paket bilan tashkil qilish mumkin. Tizimning tarmoq qurilmalari sahifani qayta ishlash jarayonini kuzatish imkoniyatiga ega - sahifa hozirda kim tomonidan ochilgan, skanerlangan, tanilgan, tekshirilgan va hokazo. Bitta foydalanuvchi tomonidan sahifaga kiritilgan o‘zgarishlar bir xil paket bilan ishlaydigan barchaga ko‘rinadi[13].

Hamkorlik imkoniyatlariga CuneiForm 2000 Master tizimi[14] ham ega bo‘lib, u CuneiForm 2000 muhitining o‘zi va matn muharririga qo‘shimcha ravishda ommaviy skanerlash va tanib olish uchun o‘rnatilgan dasturiy ta’minot blokini, shuningdek, skanerlardan foydalanish uchun dasturiy vositalarni o‘z ichiga oladi. mahalliy tarmoq.

Matnli hujjatlarni mantiqiy bog‘lanishi ko‘p yillik tadqiqot va o‘rganish natijalariga asoslangan. Bu algoritmlar ma’lumotlar analizi, matematik, statistika, yadroviy ko‘chirish, tarjima, tarix, ta’lim va boshqa sohalardan foydalanib yaratilgan[15].

Xulosa.

Ushbu tadqiqot ishining o‘rganilganligi bir necha bosqichlarda ko‘rsatiladi. Bunday algoritmlar ko‘pincha quyidagi bosqichlardan o‘tkaziladi:

Ma’lumotlarni tahlil etish: Bu bosqichda, matnli hujjatlardagi ma’lumotlar (so‘zlar, jumlalar, paragraflar va hokazo) avtomatik ravishda tahlil qilinishi kerak. Bu tahlil jarayonida, so‘zlar, jumlalar va boshqa elementlar belgilanadi va ularning bog‘liqliklarini aniqlash uchun statistik analiz qilinishi mumkin.

Bog‘liqlik tahlili: Ma’lumotlar tahlili jarayonida belgilangan elementlar orasidagi bog‘liqlik ko‘rsatkichlarini aniqlash uchun, bog‘liqlik tahlilini o‘tkazish lozim. Bu tahlil jarayonida, so‘zlar, jumlalar va hokazo orasidagi bog‘liqlik darajalari tahlil qilinishi kerak.

Bog‘liqlik matritsasi yaratish: Elementlarning bog‘liqlik darajalari tahlil qilinganidan so‘ng, bog‘liqlik matritsasi yaratiladi. Bu matritsa elementlar orasidagi bog‘liqlik ko‘rsatkichlarini ko‘rsatadi va mantiqiy bog‘lanishlarni aniqlash uchun kerakli algoritmlar yaratishga imkon beradi.

Algoritmlarni ishga tushirish: Bog‘liqlik matritsasi yaratilgandan so‘ng, bog‘liqlikning mantiqiy bog‘lanishlarini aniqlash uchun, kerakli algoritmlar ishga tushiriladi. Ko‘pincha, bu algoritmlar statistik, ma’lumotlar analizi va mashinali o‘qitish asosida ishlaydi.

Natijalarni tahlil qilish: Algoritmlar ishga tushirilgandan so‘ng, natijalar tahlil qilinadi va to‘g‘ri yoki noto‘g‘ri javoblar aniqlanadi. Bunday javoblar ma’lumotlar analizi, tartiblash, kategoriya yaratish va boshqa metodlar yordamida tahlil qilinishi mumkin.

Foydalanilgan adabiyotlar:

1. Choryorqulov G‘.H., & Qosimov N.S. (2023). ELEKTRON JADVAL MODELINING TAVSIFLANISHI. PEDAGOGS Jurnali, 30(3), 67–73.

2.TA’LIMDA DASTURLASH JARAYONINI BAHOLASHGA ASOSLANGAN AVTOMATLASHTIRILGAN TIZIMNI TADBIQ ETISH N Nizomiddin International Journal of Contemporary Scientific and Technical Research, 24-28

3. Роль анализа текстовых связей в электронных документах в информационной безопасности Г Чорркулов, Н Норматов, А Мамараимов Информатика и инженерные технологии 1 (1), 67-71

4. Tanib olish modullarini dasturiy amalga oshirish А Мамараимов, Г Чорёркулов, Н Норматов Информатика и инженерные технологии 1 (2), 38-44

5. Ta’lim tizimida baholash tizimini avtomatlashtirishni joriy etish jarayonlari va foydalanish metodlari Н Норматов, А Мамараимов Информатика и инженерные технологии 1 (2), 356-359

6. Amrullayevich K. A., Obid o'g'li S. J. ELEKTRON TALIM MUHITIDA TALABALARDA AXBOROT BILAN ISHLASH KOMPETENTLIKNI SHAKLLANTIRISH //International Journal of Contemporary Scientific and Technical Research. – 2022. – С. 641-645.

7. Obid o’g A. S. J. et al. Numpy Library Capabilities. Vectorized Calculation In Numpy Va Type Of Information //Eurasian Research Bulletin. – 2022. – Т. 15. – С. 132-137.

8. Javlon X. et al. Классификатор движения рук с использованием биомиметического распознавания образов с помощью сверточных нейронных сетей с методом динамического порога для извлечения движения с использованием датчиков EF //Journal of new century innovations. – 2022. – Т. 19. – №. 6. – С. 352-357.

9. Obid o'g'li S. J., Nodir o'g'li X. A., Jasurjonovich B. J. SUPERVISED LEARNING REGRESSION ALGORITHM SIMPLE LINEAR REGRESSION //Academia Science Repository. – 2023. – Т. 4. – №. 04. – С. 69-76.

10. Javlon, Kholmatov, and Mustafoyev Erali. "STRUCTURE AND PRINCIPLE OF OPERATION OF FULLY CONNECTED NEURAL NETWORKS." International Journal of Contemporary Scientific and Technical Research (2023): 136-141.

11.Юсупович Ҳ. Ж., Эргашев С. Б. Ў. МАКТАБ ЎҚУВЧИЛАРИДА АХБОРОТ БИЛАН ИШЛАШ КОМПЕТЕНЦИЯСИНИ РИВОЖЛАНТИРИШИ МОДЕЛИ //JOURNAL OF INNOVATIONS IN SCIENTIFIC AND EDUCATIONAL RESEARCH. – 2022. – Т. 2. – №. 13. – С. 189-194.

12. Kholmatov Javlon, & Mustafoyev Erali. (2023). STRUCTURE AND PRINCIPLE OF OPERATION OF FULLY CONNECTED NEURAL NETWORKS. International Journal of Contemporary Scientific and Technical Research, 136–141.

13. Mixliyevich Y. R., Abdullaevich U. X., Ibroxim o'g'li N. A. TALABALARDA ALGORITMIK KOMPETENTLIKNI RIVOJLANTIRISHDA UMUMKASBIY FANLARNING IMKONIYATLARIDAN FOYDALANISH HAQIDA //International Journal of Contemporary Scientific and Technical Research. – 2022. – С. 715-717.

14. Umarov X. PEDAGOGICAL CONDITIONS OF DEVELOPMENT OF PROGRAMMING COMPETENCE IN FUTURE ENGINEER PROGRAMMERS //International Journal of Contemporary Scientific and Technical Research. – 2023. – №. Special Issue. – С. 155-161.

15. Abdullaevich U. X. TOPSHIRIQLAR TALABALARDA ALGORITMIK KOMPETENTLIKNI RIVOJLANTIRISHNING METODIK TA'MINOTI SIFATIDA //International Journal of Contemporary Scientific and Technical Research. – 2022. – С. 679-682.