ӘОЖ 004.9

БАНК АҚПАРАТТЫҚ ЖҮЙЕЛЕРІНДЕГІ ҚҰЖАТТАРДЫ ТАНУ ВЕБ-ПЛАТФОРМАСЫ

Төрекұл Ернұр
6В06118 - «Ақпараттық жүйелер және технологиялар» ББББ 4 курс студенті
Қалмахамбет Нұрәділ Қанатұлы
6В06120 - «Бағдарламалық инженерия» ББББ 3 курс студенті
М.Х.Дулати атындағы Тараз университеті, Тараз қ., Қазақстан Республикасы
Ғылыми жетекші Адилова А.К.
ak.adilova@dulaty.kz

Қазіргі қаржы секторы, әсіресе Қазақстан сияқты дамып келе жатқан нарықтарда, құрылымдалмаған банктік құжаттарды (мысалы, түбіртектер, төлем тапсырмалары, транзакция үзінділері) тиімді өңдеуде айтарлықтай қиындықтарға тап болады. Дәстүрлі әдістер қолмен деректерді енгізуге немесе жалпылама OCR (оптикалық таңбаларды тану) шешімдеріне қатты сүйенеді, ал бұлар әртүрлі, сапасы төмен құжаттармен және жергілікті тіл нұсқаларымен (орыс/қазақ терминологиясы) жұмыс істегенде дәлдіктің төмендігінен зардап шегеді.

Банк ақпараттық жүйелеріндегі құжаттарды тану және өңдеу тиімділігін арттыру веб платформасы (MVP) осы олқылықтың орнын толтырады, өйткені ол ірі қазақстандық банктердің бірегей құжат форматтары мен қаржылық терминологиясына арнайы бейімделген автоматтандырылған, жоғары дәлдіктегі шешімді ұсынады.

Бұл жұмыс осы мәселені шешу үшін заманауи жасанды интеллект әдістерін біріктіретін интеллектуалды құжаттарды өңдеу жүйесін ұсынады. Платформаның негізіне PaddleOCR сияқты озық OCR технологиялары және Llama 3 сияқты үлкен тілдік модельдер (LLM) алынды. OCR кескіннен "шикі" мәтінді шығарып алса, LLM оны контексттік тұрғыдан түсінеді, қателерін түзетеді және Kaspi, BCC, Halyk сияқты әртүрлі банк форматтарынан негізгі деректерді (сома, күн, тараптар) бірыңғай JSON құрылымына келтіреді. Бұл тәсілдің басты артықшылығы - тек дәлдікті арттырып қана қоймай, Ollama арқылы жергілікті өңдеуді қамтамасыз ету, бұл құпия қаржылық деректердің қауіпсіздігін сақтайды. Осылайша, бұл зерттеу қолмен енгізуді толығымен автоматтандыруға және операциялық тиімділікті арттыруға бағытталған.

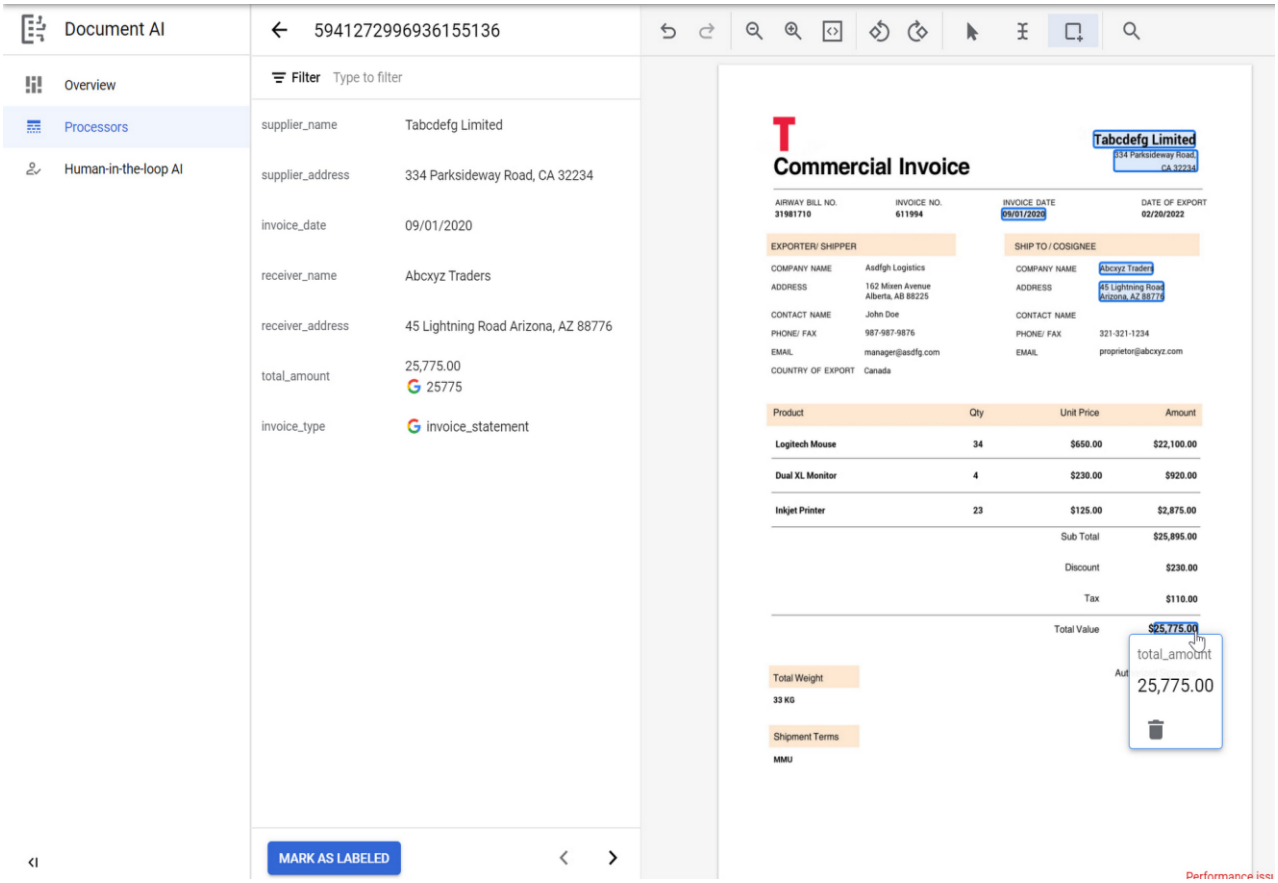
Бұл жұмыстың өзектілігі келесі факторлармен негізделеді: операциялық шығындарды азайту: құжаттардың үлкен көлемін қолмен өңдеу, уақытты көп қажет ететін, қателік ықтималдығы жоғары процесс болып қала береді. Ұсынылып отырған Интеллектуалды Құжат Өңдеу (IDP) жүйесі бұл шығындарды автоматтандыру арқылы айтарлықтай төмендетуді көздейді. Технологиялық олқылықтарды жою: Қазақстан нарығына тән әртүрлі банктік үлгілер (Kaspi, Halyk, BCC) және жергілікті тіл ерекшеліктері (орыс/қазақ кириллицасы) жалпылама OCR бағдарламаларының дәлдігін айтарлықтай төмендетеді. Жұмыс осы олқылықты ЖИ-ге негізделген арнайы шешім арқылы толтыруды мақсат етеді. Деректердің тұтастығы және стандартталуы: Банктік ақпараттық жүйелердегі деректерді стандарттау және жоғары дәлдікпен JSON форматына түрлендіру - бұл аналитикалық және бухгалтерлік бағдарламалармен өзара әрекеттесу (интеграция) үшін маңызды талап болып табылады.

Нарықта жалпы мақсаттағы бірнеше құжаттарды өңдеу және интеллектуалды автоматтандыру шешімдері (аналогтар) болғанымен, платформа өзінің нақты фокусымен және архитектурасымен ерекшеленеді. Коммерциялық бұлттық шешімдер (мысалы, Google Document AI, Azure Form Recognizer)

Бұл шешімдер платформаға балама ретінде ұсынылатын жоғары деңгейлі, ауқымды және әдетте ақылы баламаларды білдіреді.

Платформамен салыстыру:

- Артықшылықтары: Коммерциялық бұлттық шешімдер жоғары ауқымдылықты және кең үлгілерді қолдауды ұсынады, көбінесе жаһандық құжат түрлерін қамтиды.
- Шектеулері: олар платформа дизайн принциптерімен салыстырғанда айтарлықтай қиындықтар туғызады. Олардың басты шектеуі - деректердің құпиялылығы мен егемендігіне қатысты үлкен мәселе, өйткені құпия қаржылық деректерді өңдеу сыртқы, жергілікті емес бұлттық ортаға жіберуді талап етеді.



1-сурет. Google Document AI интерфейсі

Ашық қайнар көзінен алынған жалпы OCR (мысалы, Tesseract)

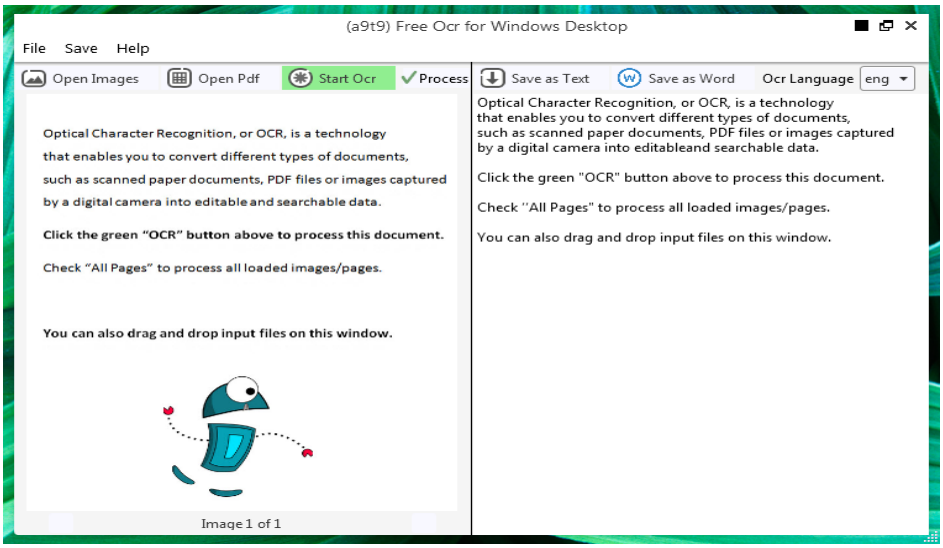
Бұл шешімдер кескіндерден қарапайым мәтінді тануды орындайтын, тегін, оңай қолжетімді кітапханалар болып табылады.

Платформа MVP-мен салыстыру:

- Артықшылықтары: Жалпы OCR шешімдерін орнату оңай және олар тегін қолданылады.
- Шектеулері: Олар жоба аясында жиі кездесетін сапасы төмен, әртүрлі түбіртектермен және күрделі банктік құжаттармен жұмыс істегенде дәлдігі төмен болады. Сондай-ақ, оларда мамандандырылған кириллица және орыс тіліндегі қаржылық терминдерді қолдау шектеулі.

Платформа MVP-нің айқын артықшылығы оның доменге тән дәлдікке және құпиялылыққа бірінші кезекте мән беретін өңдеуге (Ollama арқылы жергілікті Llama 3 моделін пайдалану) бағытталуында, бұл оны жалпыланған шетелдік шешімдерге бәсекеге

қабілетті, арнайы балама етеді. Платформа MVP-нің ең жақын аналогы – бұл арнайы жасалған шешім, бұл оның құпиялылықты-дизайн және доменге тән дәлдік арқылы жергілікті нарықтағы бәсекелік артықшылығын көрсетеді.



2-сурет. Tesseract интерфейсі

Visual Studio Code (VS Code) бұл жұмыста негізгі кодты әзірлеу ортасы ретінде пайдаланылды. Барлық Python кодтары, соның ішінде Streamlit веб-қосымшасы (app.py), PaddleOCR және Ollama логикасы бар өңдеу модульдері, сондай-ақ test.py файлдары арқылы модульдік тестілеу - осы редакторда жазылды, жөнделді және басқарылды. Оның кіріктірілген терминалы pip install -r requirements.txt арқылы тәуелділіктерді орнатуға және қосымшаны іске қосуға пайдаланылды.

Бұл жұмыста PyTorch барлық тереңдетіп оқыту операцияларының негізгі іргетасы ретінде қызмет етеді. Ол Transformers (Donut) моделі мен PaddleOCR кітапханасының жұмысы үшін қажетті күрделі тензорлық есептеулерді басқарады. Егер CUDA қолжетімді болса, PyTorch есептеулерді сонда орындайды, бұл AI модельдерінің жұмыс істеу жылдамдығын айтарлықтай арттырады.

Transformers кітапханасы құжатты "көру" арқылы түсіну үшін қолданылады. Нақтырақ айтқанда, naver-clova-ix/donut-base-finetuned-cord-v2 моделі VisionEncoderDecoderModel арқылы жүктеледі. Бұл модельдің рөлі - негізгі OCR әдістері сәтсіз болған жағдайда қосалқы әдіс ретінде жұмыс істеу. DonutProcessor суретті токенизациялап, оны модельге дайындайды, бұл жүйенің күрделі немесе нашар танылған құжаттарды да өңдеуге тырысуына мүмкіндік береді.

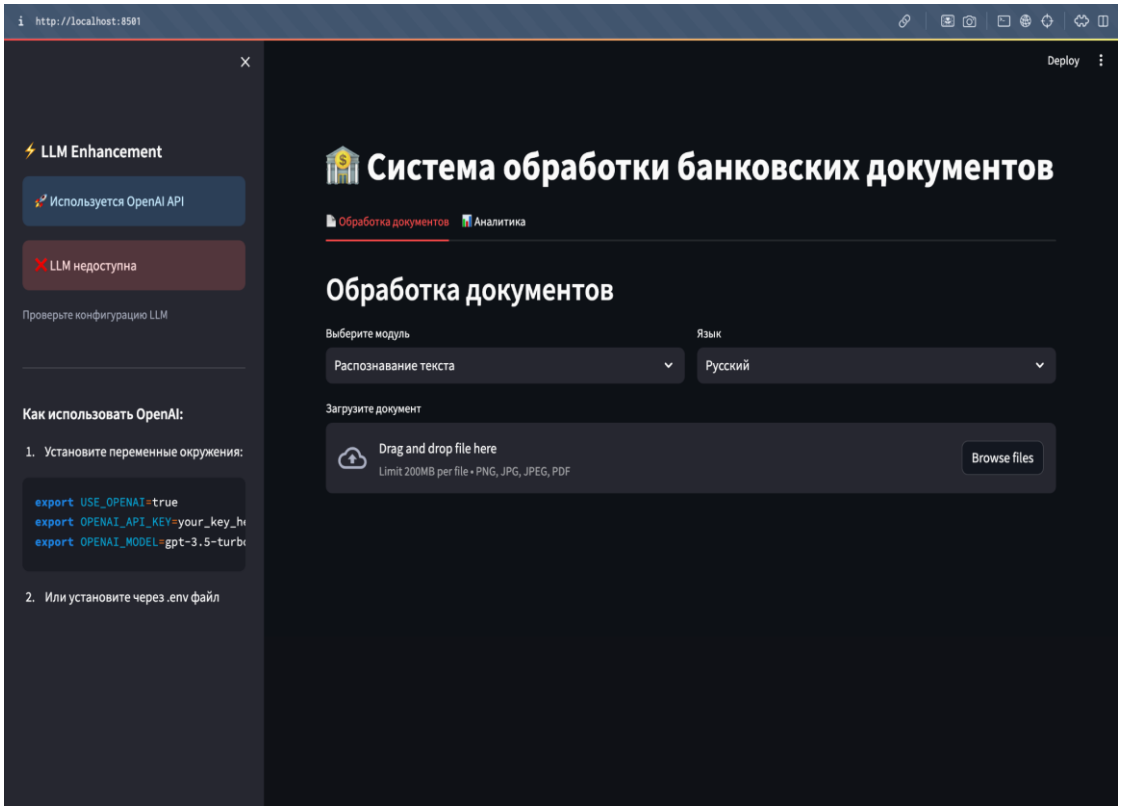
Donut моделі дәстүрлі OCR-ден айырмашылығы - ол әріптерді жеке-жеке танудан гөрі суретті біртұтас құжат ретінде талдайды. Бұл модельге құжаттың құрылымын, өрістерді, мәтін блоктарын және олардың арасындағы байланыстарды түсінуге мүмкіндік береді. Сонымен қатар, Donut алдын ала анықталған сөздіктерге тәуелді емес, сондықтан ол әртүрлі тілдердегі немесе қалыптан тыс форматтағы құжаттарда да тиімді әрекет ете алады. Модельдің тағы бір артықшылығы - шығыс нәтижесі автоматты түрде **JSON тәрізді құрылымға** келеді, бұл алынған деректерді кейінгі өңдеуде өте қолайлы етеді. Осылайша, Donut дәстүрлі OCR нәтиже бере алмаған сценарийлерде құжат мазмұнын қалпына келтіруге көмектесетін интеллектуалды резервтік әдіс қызметін атқарады.

Жұмыстағы Llama 3 моделі (llama3:8b-instruct-q4_K_M нұсқасы) Ollama арқылы жергілікті серверде іске қосылған. Оның негізгі рөлі - LLMEnhancementModule ішінде екі маңызды тапсырманы орындау: біріншіден, PaddleOCR қайтарған "шикі" мәтіндегі

қателерді контекст арқылы түзету, екіншіден, осы тазартылған мәтіннен қажетті қаржылық деректерді шығарып алып, оны құрылымдалған JSON форматына келтіру.

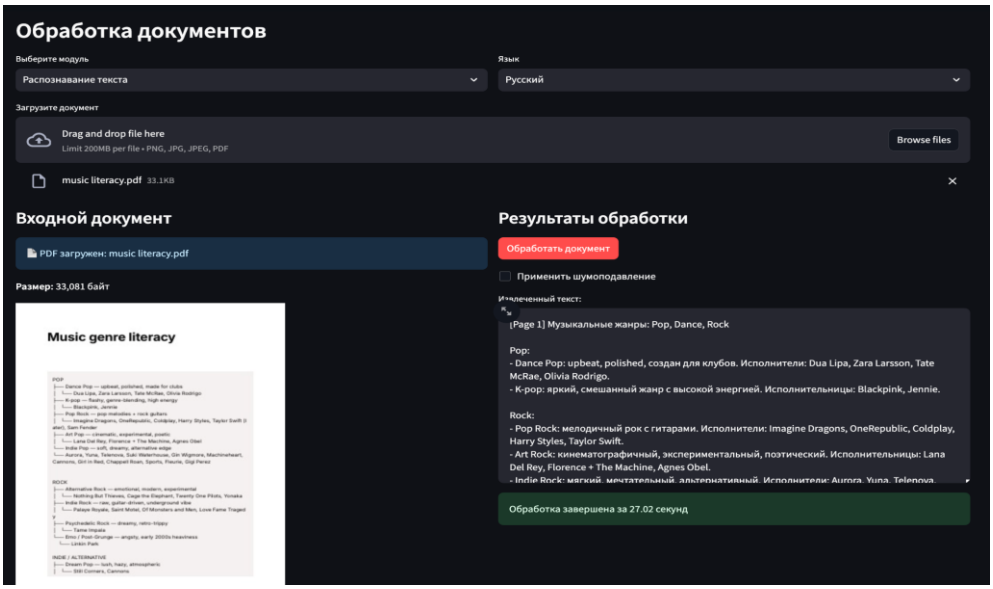
PaddleOCR және OpenCV бұл екі кітапхана бірігіп құжатты өңдеу конвейерінің басында жұмыс істейді. OpenCV бірінші кезекте кескінді алдын ала өңдеуге жауапты: ол cv2.imread() арқылы суретті жүктейді және cv2.fastColored() сияқты әдістер арқылы кескіндегі "шуды" азайтады. Осыдан кейін, дайындалған сурет PaddleOCR-ге жіберіледі. PaddleOCR кириллицаны ('ru') жоғары дәлдікпен танитын негізгі OCR қозғалтқышы ретінде ocr.ocr() әдісін қолданып, суреттен барлық мәтінді және оның сенімділік ұпайларын шығарып алады.

Streamlit бұл жұмыстың веб-қосымшасын құру үшін таңдалған негізгі фреймворк. Ол Python-native болғандықтан, күрделі AI/ML backend-імен оңай біріктіріледі және жылдам прототиптеуге мүмкіндік береді. Сонымен қатар, Streamlit сервер жағындағы есептеулерді автоматты түрде басқарып, әр өзгеріс кезінде компоненттерді қайта рендерлейді. Бұл әзірлеушіге инфрақұрылымдық мәселелермен айналыспай, AI моделінің логикасына назар аударуға мүмкіндік береді. Кэштеу механизмі ауыр ML модельдерін қайта жүктеуді тездетеді, ал кеңейтілетін архитектурасы сыртқы API-лармен немесе дерек өңдеу модульдерімен интеграциялауды жеңілдетеді. Нәтижесінде, Streamlit модельді нақты пайдаланушыларға ыңғайлы форматта көрсетуге, құжаттарды жүктеу–өңдеу–нәтиже шығару циклін интерактивті түрде ұйымдастыруға және бүкіл жүйені веб арқылы қолжетімді етуге мүмкіндік береді.



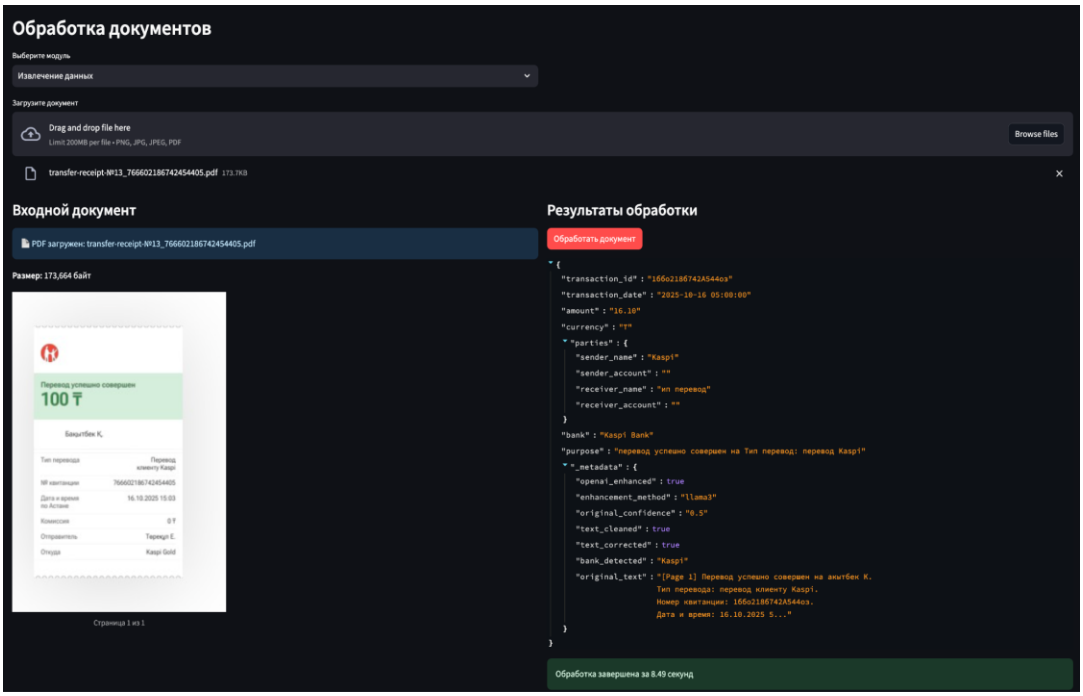
3-сурет. Платформаның жалпы интерфейсі.

Бұл негізгі интерфейс. Ол екі негізгі бетке бөлінген: "құжаттарды өңдеу" (файлдарды жүктеп, нәтиже алатын негізгі жұмыс орны) және "аналитика" (жұмыс тиімділігін көрсететін бақылау тақтасы). Сол жақтағы бүйірлік панель ("LLM enhancer") қай AI моделі (жергілікті Llama 3 немесе OpenAI) белсенді екенін көрсетеді.



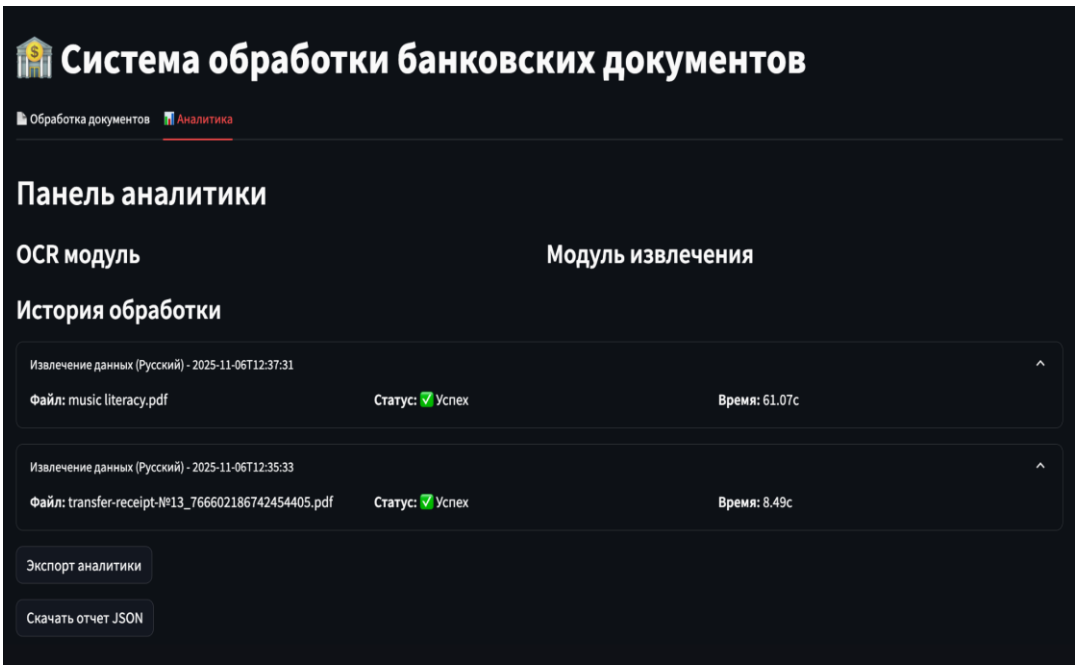
4-сурет. Платформаның мәтінді тану функциясы.

Бұл "мәтінді тану" модулі. Пайдаланушы PDF немесе JPG файлын жүктейді. Қажет болса, "шуды басу" опциясын таңдайды. Нәтижесінде, жүйе құжаттан барлық мәтінді PaddleOCR арқылы танып, оны Llama 3-пен түзетеді де, дайын, "таза" мәтінді оң жақтағы терезеде көрсетеді.



5-сурет. Платформаның деректерді JSON форматында шығару функциясы.

Бұл "деректерді шығарып алу" модулі. Бұл режимде жүйе тек мәтінді танып қана қоймай, оны талдайды. Нәтижесінде, жүйе құжаттан transaction_id, amount (сома), transaction_date (күні) және parties (тараптар) сияқты ең маңызды қаржылық деректерді тауып, оларды оң жақтағы терезеде құрылымдалған JSON форматында ұсынады.



Сурет 10. Платформаның аналитика панелі.

Бұл "аналитика" беті. Бұл бақылау тақтасы жүйенің жұмысын бақылауға арналған. Ол әр модуль үшін негізгі көрсеткіштерді көрсетеді: Сәттілік деңгейі (мысалы, ‘сәтті’), Орташа өңдеу уақыты (мысалы, 2.34с) және өңдеу тарихы (барлық операциялардың журналы). Барлық осы деректерді JSON немесе CSV файлы ретінде экспорттауға болады.

Қорытындылай келе, мақала Қазақстандық қаржы секторындағы құрылымдалмаған банктік құжаттарды өңдеудің өзекті мәселесін шешуге бағытталды. Жұмыс барысында банктік үзінділерді, төлем тапсырмаларын және түбіртектерді автоматты түрде талдап, оларды бірыңғай, машина оқи алатын JSON форматына түрлендіретін интеллектуалды құжат өңдеу (IDP) веб-платформасының минималды жұмысқа қабілетті өнімі (MVP) сәтті әзірленді.

ҚОЛДАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ:

1. М.А.Қасымбекова. Web-бағдарламалау: Оқу құралы, 2019. -219 б.
2. Б.А.Досжанова. HTML, CSS, JavaScript негіздері: Оқу құралы, 2021. -184 б.
3. Г.С.Нұрғалиева. Веб-технологиялар: Оқу құралы, 2018. -263 б.
4. А.Е.Сұлтанов. PHP және MySQL негізінде web-қосымшаларды әзірлеу. 2020. -301 б.
5. Б.Б.Байділдинов. Web-технологиялар негізінде электрондық оқулықтарды құру. 2017. -156 б.
6. Ғ.М.Мұтанов, Е.А.Есім. Жасанды интеллект жүйелері. 2019. -332 б.
7. Е.Б.Шалкеев. Python программалау тілі. 2021, -241 б.
8. Д.Т.Әбдіразақов. Машиналық оқыту негіздері. 2021, -176 б.
9. Ұ.С.Қасымова. Веб-қосымшаларды әзірлеу: деректер базасымен жұмыс. 2020, -172 б.