Szoftvertervezés és -fejlesztés specializáció

Kulcs-érték párkereső rendszer Tesseract OCR-en és utófeldolgozáson keresztül

Kaló Áron

Konzulens: Sipos Miklós

Absztrakt

Az optikai karakterfelismerő rendszerek lehetővé teszik a szöveg kinyerését képekből. Sok esetben ez elegendő lehet, de vannak olyan esetek is, amikor kulcs-érték párokra van szükség. Ebben a dolgozatban az open source Tesseract OCR rendszer alkalmazását vizsgáljuk a szöveges adatok képekből történő kinyerésére, és kulcs-érték pár keresést végzünk. A képi zaj minimalizálása szükséges képfeldolgozó algoritmusokkal a felismerés előtt. A Tesseract kimenetén utófeldolgozási eljárásokat kell alkalmazni. Ezek az utófeldolgozók átalakíthatják az OCR rendszer által végzett felismerés eredményét. Ezek javíthatják a kinyert információk pontosságát az átalakítás során, például reguláris kifejezések segítségével. A kulcs-érték pár keresés ezek után az eljárások után történik.

Kulcsszavak

Optikai karakterfelismerés, Szöveges keresés, Képfeldolgozás, Zaj eltávolítás, Reguláris kifejezések

Motiváció

A projekt célja, hogy a Tesseract OCR rendszert és utófeldolgozó algoritmusokat alkalmazva hatékony kulcs-érték pár keresést végezzen szkennelt dokumentumokban, például számlák esetén. Az alkalmazott post-processzálás és reguláris kifejezések segítségével pontosabbá válik a dokumentumból kinyert adatok kezelése, javítva a hibás felismerések korrigálását és az információ pontos kinyerését. A módszer javítja a dokumentumból kinyert adatok pontosságát, és alacsony költségű automatizálási megoldást kínál az irodai környezetek számára.

Hasonló fejlesztések

Számos kutatás foglalkozik a Tesseract OCR teljesítményének javításával és különböző verziók fejlesztésével. A legújabb verziók lehetővé teszik a párhuzamos feldolgozást és az aszinkron működést, így csökkentve a memóriahasználatot és gyorsítva a feldolgozási időt. Ezen kívül a Tesseract által kinyert szövegek post-processzálása is kiemelt kutatási téma, amely segít növelni az elért eredmények pontosságát. A post-processzálás során kulcs-érték párok keresését végezhetjük reguláris kifejezések segítségével, amelyek pontos adatkinyerést biztosítanak a dokumentumokból. Az OCR rendszerek integrációja és web API-n keresztüli adatkezelés is aktívan kutatott terület, mivel lehetővé teszi az automatizált szkennelési és adatkinyerési feladatokat más rendszerek számára, biztosítva a zökkenőmentes adatátvitelt és feldolgozást.

Sikerek

- Kari TDK 3. helyezett
- Kari Alkotói díj
- 1 konferenciacikk

Saját módszer bemutatása

A kutatás célja egy OCR rendszer kiterjesztése és teljesítményének optimalizálása volt. Az alkalmazott módszer a Tesseract OCR motort használta képekből történő szövegkinyerésre, majd a kinyert adatokat poszt-feldolgozó algoritmusokkal és szabályokkal javította. Az alábbi lépéseken keresztül zajlott a kutatás:

- **Memória alapú Tesseract integráció:** A rendszer lehetővé tette, hogy a képeket memória alapú interfészen keresztül dolgozza fel, elkerülve a fájlba mentést, amely időigényes lett volna.
- **Szöveg utófeldolgozása:** A kinyert szöveget előre meghatározott hibák alapján korrigálták, a szintaktikai hibák és a szavak környezetében előforduló hibák figyelembevételével.
- **Kulcs-érték pár keresése:** Az OCR-kimenetben található szavak alapján kulcs-érték párokat kerestek, amelyeket reguláris kifejezésekkel gyorsan és hatékonyan találtak meg.
- **Szintaktikai és szemantikai szabályok:** A rendszer lehetővé tette új szabályok generálását a tanítási folyamat során, amelyek segítettek az elismerési hibák és a szövegminőség javításában.
- **Szintetikus zaj alkalmazása:** A tanítás során szintetikus zajt alkalmaztak a digitálisan generált dokumentumok képein, hogy jobban szimulálják az optikai szkennelés során előforduló hibákat, és javítsák az OCR teljesítményét.

Eredmények értékelése

A kutatás során kifejlesztett rendszer a Tesseract OCR hatékony integrációjára épült, amely képes volt a fizikai szkennert irányítani, és jelentősen gyorsabbá tette a memória alapú fájlkezelést a hagyományos fájl interfészekhez képest. A rendszer automatikus tanulási eszközzel bővült, amely lehetővé tette az önálló tanulást és új szabályok létrehozását, így javítva a felismerési hibákat. A szintetikus zaj alkalmazása és a kulcs-érték pár keresés bevezetése szintén növelte a rendszer teljesítményét.

Az eredmények szerint a rendszer képes volt a kinyert szövegeket szintaktikai és szemantikai hibák szűrésével javítani, valamint nagyobb pontossággal felismerni a karaktereket. A jövőben érdemes lenne a képek előfeldolgozását javítani, hogy még pontosabb felismerést érjünk el.