Szakdolgozati Tématerv

Domain specifikus szöveg feldolgozása kép alapú dokumentumokon

Készíti: Marosi Márk Dániel
Gazdaságinformatikus BSc hallgató
NA0K08

A szakdolgozat témája

A digitalizáció és az automatizáció terjedésével egyre nagyobb az igény olyan programokra, melyek kép alapú dokumentumokról beolvasott szöveget képesek domain függően feldolgozni és osztályozni predikciók, szövegkörnyezet és a dokumentumon elfoglalt pozíció alapján. A szakdolgozat célja egy ilyen program elkészítése egy tetszőlegesen választott domainnel.

Funkcionalitások

Az elkészült szoftvernek képesnek kell lennie a véglegesen kiválasztott domainen a következő eljárásokat végrehajtani: a program inputja lehet egy fénykép a legelterjedtebb formátumokban (JPEG, PNG), vagy Portable Document Format (PDF). Az inputként adott fájlból első sorban egy szövegfelismerő algoritmus segítségével kinyerhető minden szöveg, és a hozzá tartozó egyéb adatok, például a képen elfoglalt pozíciója. Ahhoz, hogy a kiválasztott domainre leképezhető legyen minden kinyert adat, kelleni fog egy olyan eljárás, ami ezt az adathalmazt predikciókkal, és a szöveg tartalmának vizsgálatával fel tudja dolgozni, az adott domain tuljadonságaival rendelkező osztályra képes leképezni, majd tetszőleges formában a felhasználó rendelkezésére nyújtani, ami lehet adatbázisban való eltárolás, vagy excel, csv formátumokba való exportálás.

A témaválasztás oka

A szövegfelismerés fejlődésével egyre gyakrabban találkozunk olyan programokkal, vagy beépített funkciókkal, melyek szöveget tudnak kinyerni képekről. Ezeket a funkciókat a telefonomban is gyakran alkalmazom, előfordul, hogy az interneten is keresek ilyen web applikációkat, hogy megkönnyítsem a saját dolgom, és egy-egy szöveg saját kezűleg történő gépelését kiváltom azzal, hogy a képről kinyert szöveget kimásolom és a kívánt helyen felhasználom.
Viszont ez bizonyos esetekben további manuális korrigálást is igényel, például ha egy személyi igazolványról készült képről szeretnék minden adatot kinyerni, akkor a fentebb említett programok eredménye egy rendezetlen adathalmaz lesz, és további végfelhasználói munkát igényel az adatok megfelelő struktúrába rendezése.

A fejlesztés (tervezett) lépései

1. Jelenleg legelterjedtebb képről szöveget kinyerő programok tanulmányozása
2. Rendelkezésre álló technológiák feltérképezése
3. Domain kiválasztása
4. Implementáció
5. Tesztelés
6. Dokumentáció

A projekt végleges elkészítésének várható ideje: 2022. November 15.

A szakdolgozati tématervet készítette:
Marosi Márk Dániel, Gazdaságinformatikus (BSc) hallgató

A szakdolgozati téma elkészítésére lehetőséget biztosította:
Janurik Viktor Bálint, SZTE Szoftverfejlesztési tanszék Cisco Labor, Szeged