Rapportage

De hele geschiedenis van het probleem komt van het postbedrijf Sandd. Post verzonden door het bedrijf Sandd, passeert zijn weg door de belangrijkste felial in Appeldorn. Dit wordt beschouwd als de enige plaats waar post automatisch wordt gesorteerd en bovendien vindt sortering plaats op het niveau van postdistributie per filiaal. Al vanaf de vestigingen wordt de post handmatig gesorteerd op wijken, door middel van de transportband. Met de laatste fase van postsortering, sorteren de postbodes nu al hun post zonder een transportband. Ik besloot me te concentreren op de laatste fase van dit saaie routinewerk. Mijn hoofdtaak was om een prototype te maken van een algoritme dat het exacte adres zou kunnen bepalen van de brief(en) die wordt(en) verzonden.

De vereisten waren als volgt:

1. Algoritme slaat adres van geadresseerde en niet retouradres.
2. Pak alleen uit wat nuttig kan zijn voor toekomstige sortering.

Het werk aan de code die de afgelopen twee weken is gemaakt, bestond uit het bouwen van niet alleen het algoritme zelf, maar ook de voorbewerking van dit algoritme.De voorbewerking was niet bedoeld om ervoor te zorgen dat het algoritme alleen een voorbereide tekst ontving die de tekst in de brief imiteerde, maar de tekst die werd verkregen door de scan (foto's) van de brief en was verwerkt door een van de OCR-algoritmen. OCR "Tesseract" algoritme werd geselecteerd op de aanbeveling van een van de docenten.

Bij het testen van "Tesseract" moest ik een aantal problemen onder ogen zien. De eerste letters die ik nam voor de test werden slecht genoeg herkend. Er is een poging gedaan om dit probleem op te lossen door de achtergrondruis van elk van de foto's te verwijderen. Ik heb dit idee opgepikt uit een Adrian Rosebrock-artikel. Dit loste het probleem niet op vanwege het feit dat deze functie al in Tesseract was ingebouwd. Toen werd besloten om naar de foto's van een betere kwaliteit te gaan. Zoals later bleek, had dit een aanzienlijke invloed op het resultaat. De tekst werd beter herkend. Ook de nauwkeurigheid van de foto's heeft een grote invloed op het resultaat.

Naast de toename van de kwaliteit van foto's, is de tijd van het proces van het extraheren van tekst ook toegenomen.. De reden hiervoor was niet alleen de analyse van een groot aantal pixels, maar ook het aantal herkenbare woorden (karakters). Om dit probleem gedeeltelijk op te lossen, werd besloten de foto's bij te snijden. Het werk van de code was aanzienlijk versneld, maar er werd ook vastgesteld dat het niet de moeite waard was om bijsnijdfoto's te misbruiken. Toen de foto aanzienlijk werd bijgesneden, herkende het OCR-algoritme niets meer.

Het idee van mijn algoritme was dat het ophalen van een adres die bestond uit een postcode en een huisnummer, omdat ik dit voldoende informatie vond voor een toekomstige sortering.

Werking van algoritme werd als volgt gedaan:

In de tekst wordt teken voor teken gecontroleerd op de aanwezigheid van vier cijfers na elkaar (als de gevonden cijfers overeenkomen met de eerder voorbereide postcoden van ontvangers, dan wordt deze opgeslagen)

Naast de vier gevonden cijfers worden de resterende twee letters van de postcode toegewezen.

Nadat een volledige postcode is gevonden, behoudt de code de combinatie van tekens die voor de postcode staan, wat het huisnummer blijkt te zijn.

Het adres wordt opgeslagen en het algoritme blijft zoeken naar de aanwezigheid van de resterende postcoden.

Als een andere postcode werd gevonden in de test met dezelfde combinatie van getallen, krijgt de letter het label "overslaan". (ook het label "overslaan" krijgt een letter waarin geen enkele postcode is gevonden)

De laatste stap is het controleren en verwijderen van onnodige gegevens in de adressen. Bijvoorbeeld, zoals de aanwezigheid tussen het huisnummer en de postcode, labels van het land.

Een van de nadelen van het algoritme dat ik vond was dat het algoritme het retouradres niet van de geadresseerde kan onderscheiden in de aanwezigheid van postcoden die tot dezelfde wijk behoren. Het algoritme is ook gebaseerd op het standaardformaat van de beschrijving van het adres in de brief.Als om een of andere reden het huisnummer niet onmiddellijk voor de index verschijnt, maakt het algoritme een fout door een combinatie van tekens vóór de index in te voeren.

Referenties:

https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/