Universiteit van Amsterdam

Bachelor thesis - Project Plan

Document Splitting

10 november 2021

Student:
Daan Kuijper
11890177

Begeleider:
Maarten Marx

Samenvatting

Door documenten in te scannen kunnen deze gedigitaliseerd worden. Dit gaat veelal in de vorm van PDFs. Deze PDFs bestaan dan echter niet uit tekst, maar uit afbeeldingen en zijn dus niet te doorzoeken op inhoud. Het doel van dit project is om door middel van een OCR deze delen van een PDF ook op tekst te ontleden, om zo een hele collectie aan PDF documenten te kunnen doorzoeken op trefwoorden.

Inhoudsopgave

1 Introductie

Het splitsen van enorme PDFs in "echte documenten" is een probleem waar meerdere organisaties mee worstelen. Een voorbeeld hiervan zijn de gepubliceerde WOB-documenten door Rijksoverheid[1], deze bestaan voor een groot deel uit stapels achter elkaar ingescande documenten. Dit heeft als gevolg dat een groot deel van deze PDFs bestaan uit scans (ofwel afbeeldingen) waardoor het gebruik van een woordzoeker op de inhoud niet mogelijk is en orientatie per pagina uiteen kan lopen.

Het doel van dit project is om voor alle gepubliceerde WOB-documenten een zoekmachine te bouwen, waarbij het mogelijk is om op basis van zoekwoorden de gehele collectie aan PDFs te doorzoeken. Het resultaat van een zoekopdracht bestaat uit gevonden fragmenten uit de tekst of uit de PDF meta-data. Hierbij zal rekening moeten gehouden met een bepaalde rangorde in alle gevonden resultaten. De hoofdvraag die hierbij gesteld zal worden is: hoe is de collectie aan gepubliceerde WOB-documenten volledig op inhoud doorzoekbaar te maken?

2 Methode

Als start voor het project is de collectie met WOB-documenten van Rijksoverheid nodig. Dit zal door middel van de publieke API opgehaald worden. De data die van deze API

[1] COVID-19 WOB-documenten Rijksoverheid: `https://wobcovid19.rijksoverheid.nl`

opgehaald kan worden opgehaald wordt dan omgezet in een formaat wat gemakkelijk te gebruiken is voor de rest van het programma. Ieder WOB-document bestaat uit meerdere PDFs: bijvoorbeeld een besluit, inventarislijst en aanvullende documentatie. Uit deze PDFs moeten alle meta-data en tekst elementen opgehaald. Dit is de eerste informatie die gebruikt kan worden door een zoekmachine voor het vinden van een ingevoerd trefwoord. Aanvullend moet ook voor iedere PDF pagina bepaald of deze bestaat uit een ingescand element. Als dit het geval is zal een OCR deze pagina's moeten inlezen om nog aanvullende tekst te ontleden.

Het gehele project zal ontwikkeld worden in .NET 6. .NET is cross-platform en dus zal dit project zonder moeite werken op alle type besturingssystemen. Ook zou het project gemakkelijk kunnen werken op een server of smartphone. Voor .NET bestaan ook bewezen betrouwbare extensies voor zowel PDF bestanden manipuleren, evenals voor OCR functionaliteit. Tevens is voor .NET 6 gekozen omdat het recent is uitgekomen en ondersteund zal blijven tot minstens 2024.

Voor de OCR functionaliteit van dit project is gekozen voor Tesseract (Smith 2007). De Tesseract-ocr extensie voor .NET is veelal getest en bewezen stabiel in het behalen van resultaten. De tesseract extensie draait lokaal en is dus onafhankelijk van een server. Tesseract is in vergelijking met alternatieve OCRs minder sterk in bepaalde real-world scenarios (Tafti e.a. 2016). Echter voor de functionaliteit nodig voor dit project zou tesseract kwalitatief meer dan voldoende moeten zijn. Ook kunnen de prestaties van tesseract verbeterd worden door de start afbeelding te manipuleren (Sporici, Cușnir en Boiangiu 2020). Hier kan tijdens dit project naar gekeken worden waar voordelig.

2.1 Software

- **Portable Document Format**: Het portable document format (PDF) is een universeel bestandsformaat ontwikkelt door Adobe in 1992 voor het tonen van documenten bestaande uit geformatteerde tekst en afbeeldingen. De hoofdgedachte achter PDF is dat het op iedere machine de digitale documenten hetzelfde weergeeft, ongeacht hardware, besturingssysteem en applicatie software. Deze garantie van formaat behoudt maakt PDF de standaard als het aankomt op het digitaliseren van juridisch en wettelijke documenten.

- **.NET 6**: .NET is een applicatie framework ontwikkeld door Microsoft[2], voor het ontwikkelen van software voor zowel Windows, Linux als macOS. Hoewel .NET meerdere programmeertalen ondersteund, is C#[3] het meest gebruikelijk en beschikt over de meeste extensies en support. De meeste recente versie van .NET (.NET 6) is uitgekomen in november 2021. Deze nieuwste versie brengt prestatie verbetering met zich mee en garandeert ondersteuning voor de aankomende drie jaar.

- **Pdfsharp**: PDFsharp[4] is een open source .NET extensie die het mogelijk maakt om binnen C# programma's code PDF documenten te maken, lezen en manipuleren. De extensie is tevens gratis te gebruiken.

- **Tesseract**: Tesseract is een Optical Character Recognition (OCR) software en beschikbaar als extensie voor vele frameworks en programmeertalen, onder andere voor .NET. Het is open source en hierom zijn deze extensies gratis te gebruiken.

2.2 Planning

Dit project is uiteen te zetten in verschillende onderdelen. Hieronder een aantal van de punten die behandeld zullen worden, in volgorde van tijdbesteding:

[2]Microsoft's .NET framework: `https://dotnet.microsoft.com/`

[3]Microsoft's C# programmeertaal: `https://docs.microsoft.com/en-us/dotnet/csharp/tour-of-csharp/`

[4]PDFsharp documentatie: `http://www.pdfsharp.net/MainPage.ashx`

- **WOB-documenten**: Het dynamisch ophalen, documenteren en categoriseren van de gepubliceerde WOB-documenten van Rijksoverheid. Hiervoor wordt één week gerekend.
- **PDF**: Het inlezen van de WOB-document en hierbij alle betreffend PDFs ophalen. Uit deze PDFs moet meta data worden uitgelezen en opgeslagen. Uit deze meta data moet worden bepaald wat om welk type document het gaat. Ook moet de tekst in PDF gevonden en opgeslagen. Hier kunnen maximaal twee weken aan besteed.
- **Zoekmachine**: Nadat een lijst van alle PDF en hun respectievelijk meta data en inhoud is opgeslagen, moet er een zoekmachine gebouwd worden om door middel van trefwoorden hier door te kunnen zoeken.
- **OCR**: Als voor een PDF pagina bepaald is dat deze bestaat uit een ingescand deel of een afbeelding, moet een OCR ingeschakeld worden. Deze OCR haalt dat de tekst uit deze afbeelding en voegt deze toe aan de bestaande tekst van de PDF. Ook moet de OCR kunnen bepalen wat de orientatie is van de pagina.
- **Scriptie**: Gedurende het gehele project zal gewerkt worden aan de scriptie. Maar de nadruk komt erop in de laatste vier weken voor de eerste inlever datum.

Referenties

Smith, Ray (2007). "An overview of the Tesseract OCR engine". In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Deel 2. IEEE, p. 629–633.

Sporici, Dan, Elena Cușnir en Costin-Anton Boiangiu (2020). "Improving the Accuracy of Tesseract 4.0 OCR Engine Using Convolution-Based Preprocessing". In: *Symmetry* 12.5. ISSN: 2073-8994. DOI: 10.3390/sym12050715. URL: https://www.mdpi.com/2073-8994/12/5/715.

Tafti, Ahmad P e.a. (2016). "OCR as a service: an experimental evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym". In: *International Symposium on Visual Computing*. Springer, p. 735–746.