ILT-datalab; 17 mensen incl PhD & stagairs.
Er is daarnaast bij ILT een soort BI afdeling naast datalab (het Analyse team).
Datalab ILT werkt met afbeeldingen, tekst, etc.

scrape tool om html/PDF op internet te vinden.
Eerst data verzamelen. En dan kijken wat erin staat. Streamline dashboard.
Bijv. (specifieke use-case) Google knowledge & google maps scrapen: koppelen aan kvk via API. E dan kun je dus zien welke bedrijven gecertificeerd zijn. Koppelen aan advertenties (html scrapen). Advertenties: door welk bedrijf gedaan: toplevel domain (url) en die dan door google knowledge graph/maps, etc. cross referenties. Adressen in platte tekst zoeken is lastig en onbetrouwbaar, dus dat doen ze niet.

Hoeft allemaal niet 100% correct te zijn; doel is dat inspecteur makkelijk kan zien welke bedrijven hij moet controleren.

PFDFs gedownload: labelen op documentniveau en dan AI classificeren
classificatie, keyword extractie, namen, samenvatten.

Samenvatten: layout parser & detectron2; 1000 PDF's, meestal nota's kost een dag lokale laptop.

Doel is: we willen gewoon de hele inhoud in 1x hebben; ontsluiten van documenten en weten wat erin staat.
Unstructured IO → daar even kijken; alle parsers bij elkaar.
Je krijgt de titels, maar niet de cascade levels.
Hugging face → open source getrained modellen die je kunt downloaden en draaien.
Open source samenvatten: knipt de tekst zelf in stukjes en evt. Overlappen en dan samenvatten:

vergelijkbare teksten matchen: BERT (google) met embeddings.

Langchain heeft mogelijkheden om eigen chaintype te definieren. Dan kun je tellen/tegenstrijdigheden mogelijk wel detecteren.

- Samenvattingstool: layout parser, detectron2 voor structuur
- Hugging face open source model (lokaal gedraaid) voor samenvatten.
  - Layout parser wordt alleen gebruikt om PDF-→1 blob text te doen.
  - Dan alle text in 1x aan samenvattingstool aanbieden.
  - Tool snijdt dan willekeurig de tekst door voor size/token limiet.
  - Gebruikt een overlap-stukje om de samenvatting goed aan elkaar te breien.
  - Logisch; Layout-parser vertelt je wel hoe de alinea's zijn opgebouwd, maar niet welke alinea bij welke sectie en hoofdstuk hoort.
- Langchain chatbot voor inhoudelijke vragen aan ILT website.
- Langchain staat je toe een eigen chaintype te bouwen voor bijvoorbeeld telvragen.
- Koppelt vergelijkbare teksten via lokale BERT (Google)
- Scraping tool voor doorzoeken internet
  - html scraping voor advertenties van bedrijven
  - via hoofddomeinnaam (url) bedrijfsnaam extraheren
  - Koppelen aan google maps en kvk
  - Dan checken of dat bedrijf de vereiste certificaten bezit.
  - PDF downloaden en samenvatten is een aftakking van deze toepassing.
  - Binnen de PDF: classificeren op documentniveau, keywords, namen, locaties extractie.
  - Doorzoekt alle html, PDF, etc. o.b.v. vraag/keywords zoals 'airco's installeren'.
  - Gebruikt google, bing, yahoo, etc.