AI-trainingset Tag de tekst voor Named Entity Recognition (NER)

De AI-trainingset voor NER is in 2020 gemaakt door de circa 150 vrijwilligers van crowdsourcingsproject Tag de tekst op Velehanden.nl. Persoonsnamen, locaties en tijdsaanduidingen zijn geannoteerd in al eerder ontwikkelde Ground Truth (GT) transcripties van 10.567 scans en gecontroleerd door drie ervaren super users. Een uitgebreide beschrijving van de gehanteerde definities is te vinden in de invoerinstructie van Tag de tekst. De Nederlandstalige teksten komen uit de 17e eeuw tot en met de 19e eeuw. Het gaat om notariële teksten uit Amsterdam, Haarlem en uit zeven andere provincies en archieven van de Verenigde Oost-Indische Compagnie (VOC). Ze zijn afkomstig uit het Stadsarchief Amsterdam, het Nationaal Archief, het Noord-Hollands Archief, en zeven andere Regionaal Historische Centra, beschreven in onderstaande tabel. De AI-trainingset is ontwikkeld i.h.k.v. de projecten De IJsberg zichtbaar maken (zoekintranscripties.nl) en Slimmer zoeken in archieven (archieveninbeeld.nl).

Archief	Naam	Pagina's	Jaartallen	Beschrijving
Nationaal Archief	VOC set 1 en 2	4439	17e en 18e eeuw	Random selectie uit VOC, 1.04.02, uit 7527-9540. GT
Noord-Hollands Archief	NHA	1403	19e eeuw	Random selectie uit oud notarieel archief Haarlem, 1617, uit 1593-1805, en nieuw notarieel archief Haarlem, 1972, uit 5-813. GT
Tresoar	RHC	64	19e eeuw	Random selectie, notarieel archief, toegangsnummer 26. GT
Gelders Archief	RHC	17	19e eeuw	Random selectie, notarieel archief, toegangsnummer 168. GT
Groninger Archieven	RHC	1	19e eeuw	Random selectie, notarieel archief, toegangsnummer 86. GT
Brabants Historisch Informatie Centrum	RHC	17	19e eeuw	Random selectie, notarieel archief, toegangsnummer 7128 (5), 7637 (12). GT
Zeeuws Archief	RHC	4	19e eeuw	Random selectie, notarieel archief, toegangsnummer: 13.2. GT
Historisch Centrum Limburg	RHC	2	19e eeuw	Random selectie, notarieel archief, toegangsnummer 09.009. GT
Het Utrechts Archief	RHC	35	19e eeuw	Random selectie, notarieel archief, toegangsnummer 34-1. GT

Collectie Overijssel	RHC	18	19e eeuw	Random selectie, notarieel archief, toegangsnummer 0122. GT
Stadsarchief Amsterdam	Inventarisnummer 1283	142	1641	Notaris Hendrik Schaef. GT
Stadsarchief Amsterdam	Inventarisnummer 2410	617	1669-1677	Notaris Jacob de Winter. GT
Stadsarchief Amsterdam	Inventarisnummer 13131	764	1749-1750	Notaris Cornelis Staal. GT
Stadsarchief Amsterdam	Inventarisnummer 13132	748	1751	Notaris Cornelis Staal. GT
Stadsarchief Amsterdam	Inventarisnummer 13133	538	1752	Notaris Cornelis Staal. GT
Stadsarchief Amsterdam	Inventarisnummer 13134	534	1753	Notaris Cornelis Staal. GT
Stadsarchief Amsterdam	Inventarisnummer 12789	597	1746	Notaris Hermanus van Heel. GT
Stadsarchief Amsterdam	Inventarisnummer 14260	627	1762-1763	Notaris Hendrik Daniel van Hoorn. GT

Beschikbaar zijn de tags en de PAGE-XML bestanden uit Tag de Tekst, waarin in de XML bestanden de door de vrijwilligers getagde labels zijn toegevoegd. De labels in deze XML bestanden zijn terug te vinden aan de hand van de vermelding van 'Locatie-aanduiding', 'Tijds-aanduiding' of 'Persoons-aanduiding'. De *offset* geeft het startpunt van de label in de regel en de *length* het aantal karakters dat hoort bij dat label. Voorbeeld:

```
<TextLine id="r114" custom="readingOrder {index:3;} Locatie-aanduiding {offset:0; length:6;} Tijds-aanduiding {offset:14; length:30;
    <Coords points="707,656 717,656 719,658 721,656 729,656 731,658 735,658 737,660 741,656 743,656 747,660 755,660 757,658 759,660
    <Baseline points="717,642 791,644 866,645 940,646 1015,647 1089,647 1164,648 1238,648 1313,648 1387,648 1462,647 1537,647 1611,6
    <TextEquiv>
        <Unicode>Velsen, opden zevenden April, agttienhonderd</Unicode>
    </TextEquiv>
</TextLine>
```

Licentie
https://creativecommons.org/licenses/by/4.0/

Credits
De AI-trainingset is ontwikkeld door de ruim 150 vrijwilligers inclusief drie controleurs (super users) op Tag de tekst, Velehanden.nl. Dit is uitgevoerd in een samenwerking met Picturae, Aincient, Sioux Technologies en Islands of Meaning en de deelnemende archieven: het Stadsarchief Amsterdam, het Noord-Hollands Archief en het Nationaal Archief (ook namens de zeven andere Regionaal Historische Centra). De al eerder gemaakte Ground Truth transcripties zijn afkomstig van deze archieven. De AI-trainingset is ontwikkeld i.h.k.v. de projecten De IJsberg zichtbaar maken (zoekintranscripties.nl) en Slimmer zoeken in archieven (archieveninbeeld.nl), mede mogelijk gemaakt door het SBIR innovatie in opdracht programma van de Rijksdienst voor Ondernemend Nederland (RVO).

Attributie (BY)