Design

Om de annotatie zo pijnloos en duidelijk mogelijk te maken zal ik proberen een simpel python script te maken dat het volgende doet:

1. Een persoon vanuit de VOC clustered dataset die een match heeft met een naam van de notaris dataset binnen een bepaalde levenshtein distance wordt gekozen. (Distance moet nog bepaald worden)
2. Informatie over deze persoon uit de VOC clustered dataset zal getoond worden zodat deze vergeleken kan worden met beschikbare informatie uit de notaris dataset. De volgende informatie zal getoond worden, naam, genormaliseerde naam, plaats van herkomst, rang, heenreis schip naam, terugreis schipnaam, datum begin van dienst, datum einde van dienst.
3. Alle notaris entries die een naam bevatten binnen de bepaalde levenshtein distance met de VOC clustered dataset naam worden geselecteerd.
4. Het dan beschikbare NER model zal over de notaris tekst heen gaan en mogelijke overeenkomsten met de VOC clustered dataset zullen worden opgeslagen. Denk hierbij aan genoemde locaties die overeenkomen met de plaats van herkomst, of woorden die binnen een bepaalde levenshtein distance vallen van de schipnamen of rang.
5. Een voor een zullen de geselecteerde notaris entries getoond worden aan de annoteur samen waarbij mogelijke matches van informatie uit de notaris entry en de VOC clustered dataset worden getoond.
6. De annoteur kan dan per getoonde notaris entry aangeven of dit wel of niet dezelfde persoon is als de persoon uit de VOC clustered dataset, waarna het script deze connectie opslaat.

Na deze stappen kunnen twee optionele stappen gemaakt worden die elkaar uitsluiten:

1. Wanneer een annoteur aangeeft dat een persoon uit de notaris dataset overeenkomt met de VOC clustered dataset worden alle volgende notaris entries van dezelfde VOC clustered persoon overgeslagen en begint de annotatie van een nieuw persoon uit de VOC clustered database.

OF

1. Wanneer een annoteur aangeeft dat een persoon uit de notaris dataset overeenkomt met de VOC clustered dataset, kan de annoteur een percentage van zekerheid aangeven. De volgende notaris entries van dezelfde VOC clustered persoon worden nogsteeds getoond. In dit geval wanneer er meerdere mogelijke matches zijn aangegeven door de annoteur kan ikzelf nog naar deze gevallen kijken.

Annotation Guidelines

Volgorde van waar naar te kijken:

1. Datum notaris document en begin of eind datum dienst, wanneer de datum van de notaris entry niet binnen een bepaald marge valt ten opzichte van de begin of eind datum dienst kan de entry al worden afgeschreven. (marge moet nog bepaald worden)
2. Wanneer er mogelijk overeenkomende informatie wordt aangegeven door het python script is het handig hier eerst naar te kijken. Volgorde van
3. Als er geen verdere informatie beschikbaar is maar de datum klopt wel kan er geprobeerd worden extra informatie te halen uit de HTR tekst

Wat te doen als er alleen een naam en datum aanwezig zijn?:

* Wanneer de 2e optionele stap uit het design wordt gehanteerd geef aan dat het een match is maar met een laag percentage van zekerheid (20-50%?)
* Anders geef aan dat het geen overeenkomst is?