InfoShare PDFtotext scanner
zones og rules opsætning

PTS softwaren behandler pdf dokumenter efter de regler, der er beskrevet i zones.txt eller rules.txt filerne i destinationsmappen. Disse regler indeholder oplysninger om ”hvad skal der ledes efter” og ”hvad skal der ske”. Samt for zones.txt filerne ”hvor skal der ledes”.

Zones og rules filerne kan redigeres med et almindeligt tekstredigeringsprogram som fx Notepad eller de kan oprettes med hjælpeprogrammet PTStools, der gør det lidt nemmere at overholde det nødvendige format og lave betingelserne.
Det generelle format for en zones linje er:
tekst####zone####betingelse####handling
Og tilsvarende for en linje i en rules fil:
tekst####betingelse####handling

Den grundlæggende funktionalitet er ”led efter en tekst med regular expressions og brug det fundne til at ændre mappenavn eller filnavn”, men derudover findes der en lang række specialfunktioner, som beskrives i det følgende.
Hvis begge filer findes, benyttes kun zones.txt
Behandling af disse filer virker **rekursivt**: Hvis en behandlet pdf fil skal flyttes i en undermappe og denne undermappe også indeholder en zones eller rules fil vil disse regler også blive evalueret.

Zonebeskrivelse:
I zones.txt filen er 2. led i hver linje en angivelse af den zone (fysisk del af et dokument), der skal kigges i. Det led findes ikke i rules.txt filens linjer, hvor der altid ledes efter betingelser i hele dokumentet.
En zone beskrivelse består af 5 tal: sidenr, x-startposition, y-startposition, bredde, højde.
X, y, bredde, højde kan findes ved at bruge systemets hjælpeværktøj selectrectangle (som kaldes automatisk, hvis man bruger PTStools).

Hvis betingelsen (indenfor zonen) er opfyldt udføres handlingen:
Generel funktionalitet:
Filnavn: Enhver handling, der slutter med +: Filnavnet ændres til teksten foran + efterfulgt af det, der blev fundet med betingelsen. Hvis flere filer i samme mappe efter denne regel skal have samme navn tilføjes automatisk et løbenummer i filnavnet.
Mappenavn: Hvis betingelsen er opfyldt ændres mappen til det navn, der står i handling. Med mindre handling indeholder ”Found value” - så ændres mappenavnet til den fundne tekst (og mappen oprettes om nødvendigt)
En par regler for sortering af alle dokumenter i separate mapper for hvert cpr nummer samt ændring af filnavn til den første fundne dato i dokumentet vil fx være:
Mappe til cpr####\b[0-3][0-9][0-1][0-9]{3}-[0-9]{4}\b####Found value
Mappe til cpr####\b[0-3][0-9][0-1][0-9]{7}\b####Found value
Datofilnavn####\b[0-3]{0,1}[0-9][-/\.]{0,1}[0-1]{0,1}[0-9][-/\.]{0,1}[1-2][0-9]{3}\b####Dato+

Ud over de generelle regler findes der en række specielle funktioner, der kan anvendes i
Funktioner vedrørende mappenavn:
m(...) betyder at det er en funktion der bestemmer mappenavn
m(**After**, '...') - vælger mappenavn til den tekst (ekskl blanke) der kommer efter den angivne tekst '…' (Fx m(After,’Fakturanr’) vil finde det der står efter teksten Fakturanr – formodentligt nummeret – og bruge det som mappenavn.

m(**FolderYear**) - vælger mappenavn til årstallet i en funden dato
m(**FolderIsoDate**) - vælger mappenavn konverteret til ISO-format (ååmmdd) fra en funden dato
m(**ForceNumeric**) - forsøger at genkende håndskrevne tal og vælger mappenavn derefter
m(**CurrentDate**, ['%formatering%']) - vælger mappenavn til dags dato, hvor '%formatering%' er en valgfri formateringsstreng hvor '%Y-%m-%d' er default. Se http://strftime.org/ for en liste over formaterings-markører
m(**LoebeNr**, ['%formatering']) - vælger mappenavn til et løbenr, som inkrementeres for hvert behandlede dokument. Formateringsstrengen er valgfri, og har '%06d' som default (dvs. at løbenr'et altid vil have 6 cifre)

Funktioner vedrørende filnavn:
f(...) betyder at det er en funktion der bestemmer filnavn
f(**After**, '...') - vælger filnavn som den tekst (ekskl blanke) der kommer efter den angivne tekst '…' (Fx f(After,'Fakturanr') vil finde det der står efter teksten Fakturanr – formodentligt nummeret – og bruge det som filnavn.
f(**FilenameYear**, '...+') - vælger filnavn til '...' + årstalsdel af funden dato, hvor '...' er et selvvalgt prefiks
f(**FilenameIsoDate**, '...+') - vælger filnavn konverteret til ISO-format (ååmmdd) fra en funden dato, hvor '...' er et selvvalgt prefiks
f(**ForceNumeric**, '...+') - forsøger at genkende håndskrevne tal og vælger filnavn derefter, hvor '...' er et selvvalgt prefiks
f(**CurrentDate**, ['%formatering%']) - vælger filnavn til dags dato, hvor '%formatering%' er en valgfri formateringsstreng hvor '%Y-%m-%d' er default. Se http://strftime.org/ for en liste over formaterings-markører
f(**LoebeNr**, ['%formatering']) - vælger filnavn til et løbenr, som inkrementeres for hvert behandlede dokument. Formateringsstrengen er valgfri, og har '%06d' som default (dvs. at løbenr'et altid vil have 6 cifre)

Funktioner, der ændrer dokumentindhold:
Anonymisering:
Remove - sværter det specificerede område, hvis regex matcher.
Stempel:
s ('<dynamisk-tilføjelse>', '<tekststørrelse>', '<tekst>') - laver et tekststempel ved de specificerede x og y koordinater.
<dynamisk-tilføjelse> kan være
- 'dd', som laver et datostempel: fx s(dd,12,"Modtaget d."),
- 'fn', som skriver filnavnet
- 'fri', som kun skriver den angivne tekst.
Stemplet fjerner ikke evt. underliggende tekst på dokumentet. Teksten må ikke indeholde , (komma).

Funktioner, der danner opsummering af indhold fra dokumenter:
csv('<indeks>', '<filnavn>') - definerer et felt i en CSV-linje med valgt indeks-nummer og filnavn.
csv-below('<indeks', '<filnavn>') - definerer et felt i en CSV-linje med tekst som kommer linjen efter den specificerede regex med valgt indeks-nummer og filnavn.
Funktionen kan fx bruges til at lave en kommaseparereret fil med alle navne og adresser fra en stak ensartede dokumenter, fx fakturaer eller ordrebekræftelser.

Generel styring af regelkontrol
Når den første regel, der definerer et mappenavn eller et filnavn er fundet ledes der ikke længere efter regler, der definerer mappe- og filnavne. Den første regel, der giver et "hit" er altså den, der bestemmer.

Styring af betinget regelkontrol:
select(<regelnavn>) - Hvis regex matcher, ignorerer programmet alle regler som ikke hedder <regelnavn> (i det ellers frie 1. led til regelnavn)
Fx anonymisering af forskellige områder i forskellige formularer:
Formular A107####1,50,50,800,200####A107####select('a107')
Formular A109####1,50,50,800,200####A109####select('a109')
a107####1,100,340,800,80####Remove
a109####1,100,520,800,80####Remove

Eksempel på regelfil:
Læsning af RegEx delen: RegEx delen kan dannes mere brugervenligt med vores PTStools program. Ved læsning af nedenstående eksempel (og håndredigering) er der et par ting, der kan være en hjælp:
\b betyder "ordgrænse" - det kan være mellemrum, linjeskift, punktum eller lignende.
(Ii) betyder "stort I eller lille i"
(staveform1|staveform2|staveform3) betyder "en af staveformerne mellem | (pipe).
[0-9] betyder et af tegnene i rækkefølgen 0 til 9 (altså "et tal")
{2} betyder det foranstående skal gentages 2 gange, {2,4} betyder mindst 2, højest 4 gange.
Der er en mere udførlig vejledning i Regular expressions quick reference.

```
01F Company1####\b(Ii)nfo(Ss)hare\b####InfoShare
02F Company2####\b(nyborg|Nyborg|NYBORG)\b####Nyborg
03F Person1a####\bArne Skov\b####Privat
04F Person1b####\bARNE SKOV\b####Privat
05F Person1c####\bArne\b####Privat
06F Person1d####\bARNE\b####Privat
07F Person2a####\b(?ix)Anton Bach Carlsen(?-ix)\b####Anton
08F Person2b####\b(?ix)Anton Carlsen(?-ix)\b####Anton
09F Money1####\b[0-9]{1,3}[.][0-9]{3}\b####Accounting
10F Money2####\b[0-9]{1,3},[0-9]{2}\b####Accounting
11F Carreg1a####\b\w(2)[0-9]{5}\b####Found value
12F Carreg1b####\b\w(2) [0-9]{5}\b####Found value
13F Carreg1c####\b\w(2) [0-9]{2} [0-9]{3}\b####Found value
14F Personid1####\b(?x)\d{10}(?-x)\b####Found value
15F Personid2####\b(?x)\d{6}-\d{4}(?-x)\b####Found value
16F Personid3####\b[0123]\d[01]\d{3}[-\s]{0,1}\d{4}\b####Found value
17N Date1####\b[0-9]{1,2}[-/.][0-9]{1,2}[-/.][0-9]{2}\b####Date-+
18N Date2####\b[0-9]{1,2}[-/.][0-9]{1,2}[-/.][0-9]{4}\b####Date-+
19N Date3####\b[0-9]{1,2}[-/.][0-9]{1,2}[-/.][0-9]{4}\b####Date-+
20N Date4####\b[0-9]{2}[-/.][0-9]{1,2}[-/.][0-9]{4}\b####Date-+
21N Month1####\b(?i)Januar(?-i) [0-9]{2}\b####Mth-+
22N Month1####\b(?i)January(?-i) [0-9]{4}\b####Mth-+
23N Month2####\b(?i)Februar(?-i) [0-9]{2}\b####Mth-+
24N Month2####\b(?i)February(?-i) [0-9]{4}\b####Mth-+
25N Month3####\b(?i)Marts(?-i) [0-9]{2}\b####Mth-+
26N Month4####\b(?i)April(?-i) [0-9]{4}\b####Mth-+
27N Month5####\b(?i)Maj(?-i) [0-9]{2}\b####Mth-+
28N Month5####\b(?i)May(?-i) [0-9]{4}\b####Mth-+
```

På InfoShares hjemmeside (http://infoshare.dk/pdf-tekst-scanner-regler/) kan du finde eksempler på andre regelfiler, blandt andet en fil med alle danske bynavne.

InfoShare, November 2019