Swedish daily newspapers data parser

Table of contents

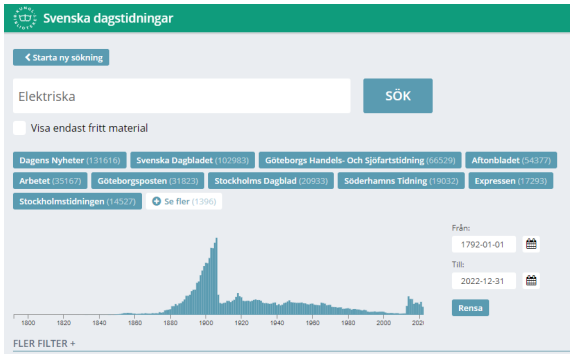
Purpose

I want to get information about the jobs advertised in Sweden to track how descriptions for jobs changed over time. I was inspired by the paper *Running out of time: using job ads to analyse the demand for messengers in the twentieth century* by Håkansson, Karlsson and La Mela that focusses specifically on messengers, and uses data from the *Svenska dagstidningar* database described by Karlsson here.

Luckily, a number of the newspapers are digitized by the National Library of Sweden.

Digitized newspapers

The repository of digitized newspapers is accessible in two ways. First, you can use a graphical user interface to search for specific terms, as shown in the image below. You can filter by publication date and newspaper.



The advantage of this way of accessing the material is that it is easy to find out if the kind of information you are searching for is accessible.

The downside is that collecting this information into a usable format is difficult. You can email yourself an image of the page of the newspaper, but it is not machine readable in this format.

The alternative way of accessing the data is to use the API created by the National Library.

The remainder of this document shows the process of accessing the data this way.

How to get the XML data from the OCR done on the newspapers

The url looks like:

"https://data.kb.se/dark-99732/bib4345612_18730208_0_32_0002_alto.xml"

You can parse the xml file the the **XML** package in R.

This is what the data looks like. It has a series of elements called `ComposedBlock` which store the data from each page of the newspaper.

```
# A tibble: 57 x 4
   name          TextBlock        .attrs    Illustration
   <chr>         <list>           <list>    <list>
 1 ComposedBlock <named list [2]> <chr [2]> <NULL>
 2 ComposedBlock <named list [2]> <chr [2]> <chr [5]>
 3 ComposedBlock <named list [2]> <chr [2]> <NULL>
 4 ComposedBlock <named list [2]> <chr [2]> <NULL>
 5 ComposedBlock <named list [3]> <chr [2]> <NULL>
 6 ComposedBlock <named list [2]> <chr [2]> <NULL>
 7 ComposedBlock <named list [4]> <chr [2]> <NULL>
 8 ComposedBlock <named list [2]> <chr [2]> <NULL>
 9 ComposedBlock <named list [3]> <chr [2]> <NULL>
10 ComposedBlock <named list [2]> <chr [2]> <NULL>
# ... with 47 more rows
```

The purpose here is to take the XML and make it into text, in the right order such that each article follows on from the previous.

It works well on this page

It returns a dataframe with the location of each word on the page, as well as a word confidence score, for how certain the OCR process is about the content of a particular word.

```
# A tibble: 1,210 x 5
    VPOS WIDTH HEIGHT CONTENT          WC
   <dbl> <dbl>  <dbl> <chr>            <chr>
 1  2038    53     33 Ett              0.73
 2  2036   312     46 krigareäfventyr. 0.81
 3  2079   229     26 dkihädl          0.44
 4  2047    19     35 y                1.00
 5  2083    57     24 Den              0.88
 6  2081    65     25 enda             0.29
 7  2079   197     32 krigshändelse    0.60
 8  2091    54     18 som              0.28
 9  2085    70     25 ännu             0.34
10  2085    69     26 varit            0.41
# ... with 1,200 more rows
```

What to do next??

Function to Visualize positions of blocks on page.

Scraper or use API to get the files

Check that complex page with adverts works.

API

Works see here

To get the files if you know the package ID and the file_name - can probably get these with the search API.

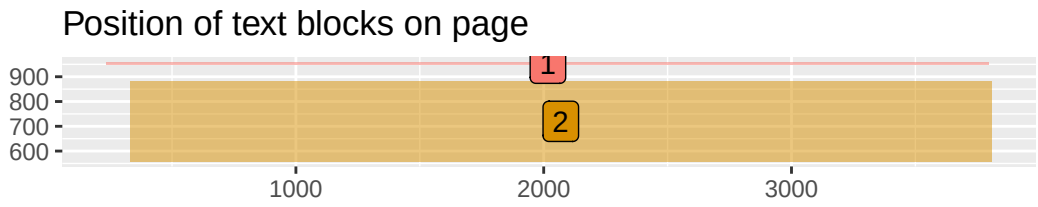
Does it work with complicated pages?? Let's try

Show 10 entries Search:

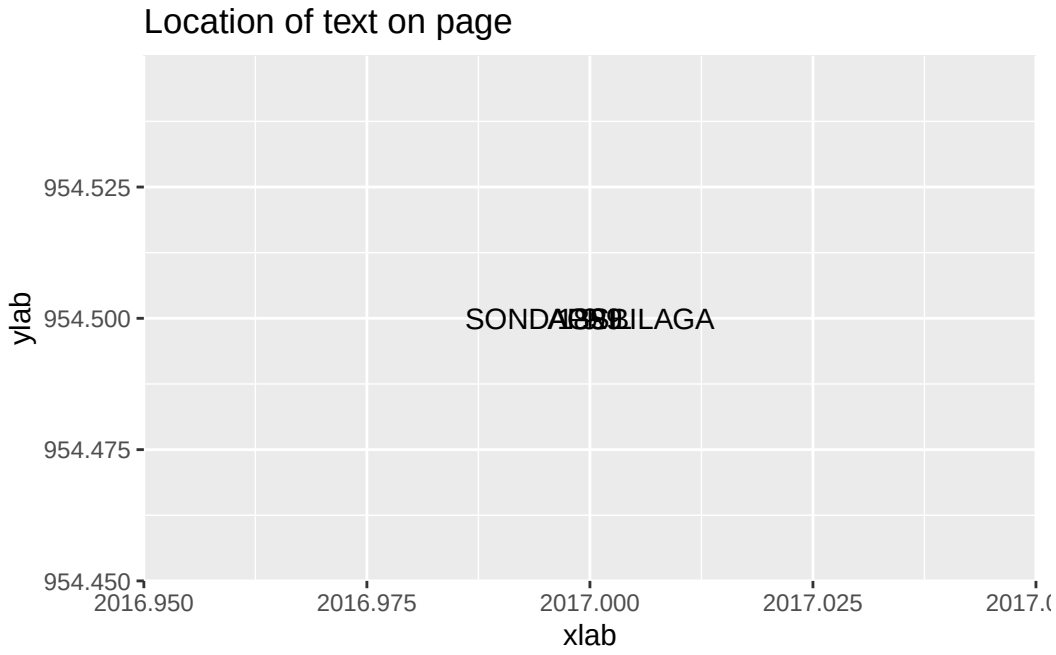
Text block	Text
1	9 APRIL SONDAGSBILAGA 1899
2	SVENSKA DAGBLADET jFör Stockholm och Landsorten Stockholm Svenskil Dagbladots tryckeri
3	Yattenhyaciiiteii i Florida Sedan några år tillbaka liar den i Floridas vattendrag och särskildt i S :t Johnsfloden förekommande vatten hya - cinten (Piaropus 1 Eichhvnia erassi - pes förökat sig i så oerhörd grad att den lägger allvarliga hinder i vägen för sjöfarten Förhållandet har i själf - va verket vållat så stora olägenheter att det till och med varit föremål för kongressens uppmärksamhet och För - enta Staternas åkerbruksdepartement har genom sin botaniska afdelning äg - nat detsamma en ingående utredning hvarur följande upplysningar torde va - ra af allmänt intresse
4	Vatten hyacinten tillhör fam Ponte - deiiaccw och har sitt hemland i Sydame- rikas tropiska och subtropiska trakter Uppmärksammad och omtyckt för sina talrika ljusblå eller violetta blommor har den blifvit föremål för vidsträckt odling i Europa och Norra Amerika och särskildt i Florida har den så full - ständigt naturaliserats att den upp - träder massvis i insjöarna och i sådana floder och åar hvilkas vatten är upp - blandadt med organiska ämnen och här och hvar bilda stagnerande sam - lingar Växten förekommer flytande i vattenytan och har endast på grundare ställen sina rötter fastade i botten Bla - dien sitta i en rosett af en till två fots höjd Bladstjälkarna äro uppsvullna och fyllda med luft och underlätta där - igenom plantans flytande Rötterna bil - da en tät kvastlik knippa al omkring två fots längd
5	öfverstelojtnant BRATJNERHJEEM unerhjelm är dateradt den 20 november 1895 X bref från London af den 7 Juli 1897 medde - lade ing O öfverstelöjtn B att han hade utarbetat sin princip äfven för styrning vid dagsljus Sedan den förre i bref den 27 sam - ma månad försäkrat den senare att Mr J P Armstrong i London hos hvilken O sedan 1 arbetat på andra uppfinningar icke hade någon del i torpedstyrningsapparaten samt att han utfört alla experiment under sin fri - tid och med egna medel var det som hr B åtog sig saken I slutet af februari 1898 kom ing O till Stockholm från London men re - ste åter till England och Amerika under april maj och juni för att studera vissa de - taljer Under september och oktober utförde Orling och Braunerhjelm med biträde af «» ångslup från k flottan på Lilla Värtan vissa förberedande experiment för att utröna dB elektriska strålarnas inverkan på den härfö konstruerade mottagningsapparaten samt for åstadkommande af rodrets vridning styiv bord eller babord Dessa försök förevisades i midten af oktober inför medlemmar af def sTftdlkat som nu är ägare at uppfinnlngm
6	Siijrbar Torped upptäckte att motståndet 1 dylika med me - tallfilspån fyllda rör minskas betydligt om I närheten försiggå elektriska urladdningar Engelsmannen Lodge förordade det Branly - ska röret såsom en särskildt känslig angif - vare af elektriska oscillationer Ännu kunde emellertid det Branlyska röret ej komma till någon praktisk användning emedan de elek - triska oscillationerna sedan de träffat de svårledande filspånen i röret — där det sitter i den elektriska strömledningen — och gjort dem lätt ledande på sam - ma gång sintrade ihop dem så att de ej efter

Visualize blocks on page

Here is where the blocks fit on the newspaper page:



Here is where the text fits on the page.


Location of text on page
954.525
954.500
954.475
954.450
ylab
2016.950
2016.975
2017.000
2017.025
xlab