Swedish daily newspapers data parser

Table of contents

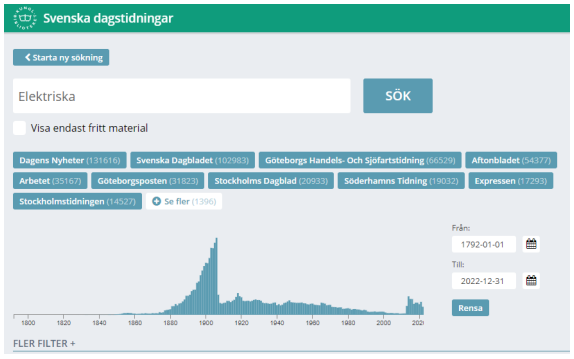
Purpose

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 Royal 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"

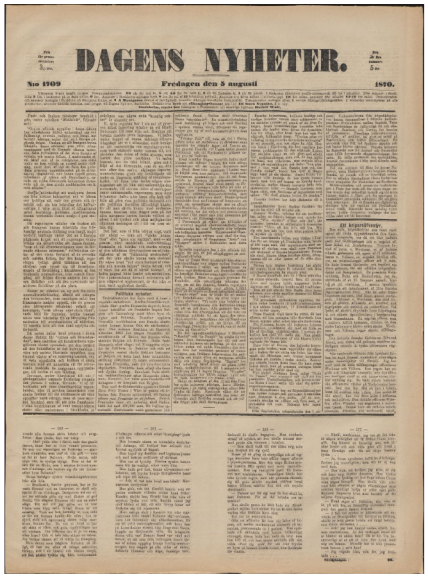
You can parse the xml file the the **XML** package in R.

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

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

Text	
blo	Text

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

Text
blo Text

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 oscillationernas upphöran - de återtogo sin förra egenskap utan i stäl - let bibehöllo sin nyförvärfvade egenskap att lätt leda strömmen För att det Branlyska röret därefter ånyo skulle bli svårledande för elektrisk ström måste det erhålla en knuff som skakade om filspånen i röret och upp - häfde lhopsintringen Detta föranledde ita - lienaren Harconi att med röret förena en hammare som automatiskt slog ned på rö - ret för hvarje gång sam detta belysts af de oscilieraaiie elektriska vågorna ten i oerhörda massor vid sjöarnas och flodernas stränder bildande täta och genom refvor och bistjälkar samman - hängande mattor hvilkas fägring dock icke förmår öfverskyla de allt mer och mer stigande olägenheterna af växtens förvånansvärda förmåga att föröka sig och vinna allt vidsträcktare ter - räng Med en lifskraftig arts hela hänsynslöshet har vattenhyacinten nu vidgat sin maktsfär så att den lägger högst betydliga hinder i vägen för sjö - och flodfarten Till och med ångbåtar - na kämpa med svårigheter att bana sig väg genom de blommande hyacintmas - sorna och som det visat sig att de af naturen lämnade fienderna till växten bland hvilka i främsta rummet märkes en parasiterande svampart icke förmå hämma dess fortgående ökning och spridning har nuui nu börjat utfunde- ra verksamma utrotnings- och in - skränkningsmedel på mekanisk väg Kampen mot inkräktaren är så mycket mer berättigad som han utom sina tra - fikhindrande egenskaper äfven är häl - sovådlig när de hopade döda individer - na öfvergå i förruttnelse och vid sjö - och flodstränderna bilda härdar för bakterier och jäsningsämnen Den ekonomiska nytta som vattenhyacin ten äger såsom ett särdeles godt foder för svin och äfven för nötkreatur är sålunda ej tillräcklig att uppväga dess §åväl för samfärdseln som i hygieniskt hänseende menliga egenskaper Vår afbildning visar några flodån - gare eller ångfärjor af den i Förenta Staterna vanliga typen hvilka arbeta sig fram genom hyacintbäddarna på S :t Johnsfloden

Text block	Text
7	C F Nu kunde det Branlyska röret börja att gö - ra tjänst genom att sluta och öppna en elek - trisk strömledning allt efter som det under kortare eller längre tid träffades af de “Hertzska vågorna såsom de oscillerande elektriska vågorna gemenligen kallas Den Orling-Braunerhjelmska uppfinningen är nu ett öfvergående”from sound to things från signalers sändande på afstånd till frambringandet af praktiska verkningar på afstånd bäggedera utan tråd Telegraferisgen utan tråd För en rätt uppfattning af den Orling-Bra - unerhjelmska uppfinningen är det alltså af nöden att — om också blott antydningsvis — vidröra telegraferingen utan tråd A Slaby professor vid tekniska högsko - lan i Berlin begagnar sig af följande appa - rater för sin telegrafering utan tråd Klämskrafvarna på en vanlig induktions - apparat sättas i förbindelse med tvenne mas - siva mässingskulor hvilka genom tillräck - ligt tjocka oledande ebonitplattor hållas på nödigt afstånd från hvarandra Vid induk - tionsapparatens igångsättande uppstå mellan
8	t kulorna kraftiga hvitglänsande elektriska gnistor hvilkas kraft ännu mera förhöjes om mellanrummet mellan kulorna fylles med olja som också är oledande i förhållande till den på kulorna befintliga elektriciteten En - ligt ett af italienaren Righi först användt förfarande låter Slaby den elektriska ström - men från induktorn ej ingå direkt till de stora kulorna utan ledningens poler få slu - ta i mindre kulor som befinna sig på lämp - ligt afstånd från de större hvarigenom en oscillation af elektriska gnistor uppstår äf - ven mellan dessa mindre kulor och de stör - re när strömmen öfvergår till de senare Detta var nu det väsentliga af Slabys af - sändningsapparat Vid mottagningsappara - ten spelar det förut omnämnda Branlyska röret hufvudrollen När de Hertzska vågor - na träffa filspånen i detta rör blifva dessa svårledande spån lättledande och otaliga små osynliga elektriska gnistor hoppa öfver dem emellan åstadkomma metallisk förbin - delse mellan spånen och sluta den elektriska strömmen som i sin tur drager nod ankaret i ett relais och — ett tecken göras på pag-

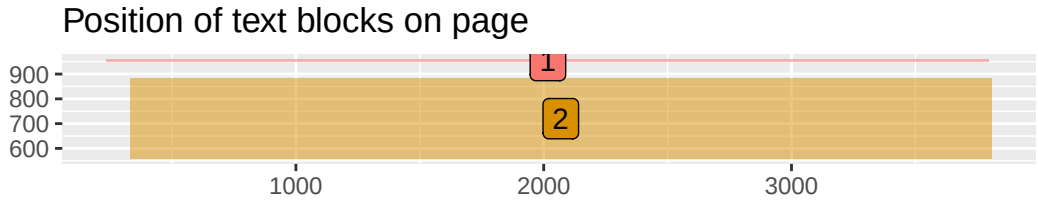
Text
blo Text

9 jAJsändninCjS Apparat Bl persremsan 1 skrlfapparaten För hvarje gång induktionsströmmen slutes på afsänd - ningsstationen uppstå sålunda tecken på mottagningsstationen Ja något mer torde vi ej behöfva säga om telegraferingen utan tråd för att läsaren skall kunna fatta det framsteg som ligger i den Orling-Braunerhjelmska uppfinningen af torpedstyrning på afstånd Denna uppfinning har ej heller fullbordats på en gång utan är tvärtom en frukt af fortlöpande uppfinningar och sedan den 23 september 1897 då den första svenska patent - ansökningen inlämnades har i Sverige dess - utom sökts inalles 9 olika kompletterande specialpatent Hvilken väsentlig olikhet des¬ sa patent företett sins emellan kan Inses däraf att först gällde det att på afstånd framkalla den styrande verkan på rodret me - dels ett knippe ljusstrålar som utkastades från afsändningsstationen mot tvenne med Selen täckta skifvor på torpedon Senare då de kraftförmedlande elektriska etervå - gorna vid försöken fått efterträda ljusvågor - na har Selenen ersatts med en kohär af tre efter hvarandra följande olika konstruktio - ner Då den senaste af dessa uppfinningar än - nu icke hunnit att medels patent skyddas i alla länder ha vi ingen rätt att här Inlåta oss på en detaljerad beskrifning af densam - ma Vi måste därför inskränka oss till föl - jande antydan om den Orling-Braunerhjelm - ska uppfinningen Den afsändnlngsapparat som uppflnnarne använda består af ett IS volt aökumulfttw - - batteri ett 25 cm gnistinduktorium en kraf - tig kondensor (30X16 cm af egen konstruk - tion och en strålkastare för Hertz-vågornas utsändande och dirigerande Apparaten för strålarnas uppfångande är af sådan konstruktion att dess känslighet kan regleras för olika afstånd Genom ett reiais och vissa mekaniska anordningar står uppfångaren i förening med styranordnin - gen för rodret Genom från afsändningsap - paraten utsända strålar åstadkommes rodret» omställning åt höger eller vänster Efter detta torde vi bli förstådda när vi an ^ifva framsteget 1 den Orling-Brauner - hjelmska uppfinningen framför Marconis och Slabys till 1 förstärkning på afsändnings - stationen af de utsända Hertz-vågornas kraft genom en så konstruerad kondensor att den medger vida större elektrisk spänning (ten - sion och i följd däraf alstrar längre och starkare gnistor som utsända kraftigare etervågor på längre håll 2 säräkild kon - struktion och användande a ,f de känsligaste ingeniören AXEL ORLINO metaller ge uppfångaren en vida större kaus lighet än de förut använda mottagningsap - paraterna ägt hvarigenom bl a nödvändig - heten af höga ledningar bortfaller 3 sam - mankopplingen af den sålunda af etervågor - na påverkade uppfångaren på torpeden med dennes styrapparat så att den “vetenskapli - ga sagan om att från land eller från annat fartyg styra en torped ute i sjön blir en sann verklighet Hvem borgar na för att apparaterna duga i Hvilka försök ha de Orling-Braunerhjelm - ska apparaterna blifvit underkastade Hvad säga vederbörande myndigheter om desam - ma Hurudana äro de financiella utsikterna för uppfinningens förverkligande Hvad beträffar den första frågan kunna vi meddela följande hvarvid vi äfven återgifva några data ur uppfinningens historia Första kontraktet om uppfinningen mellan ingeniör Orling och överstelöjtnanten Bra-

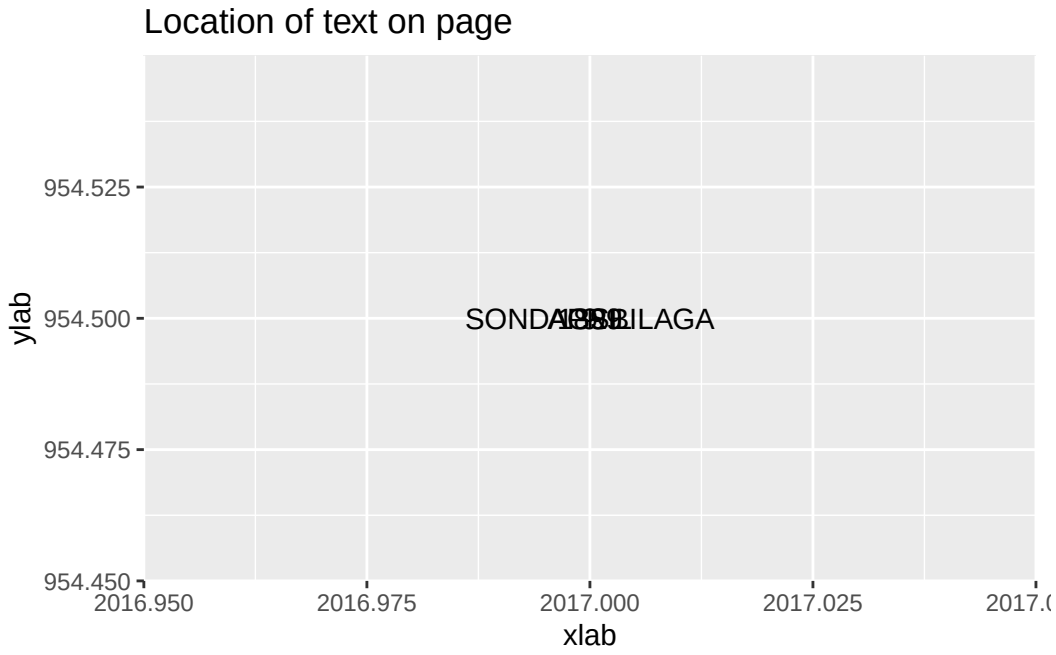
Text block	Text
10	I Florida förekommer vattenhyacin- Den Orling-Braunerhjelmska torpedstyrningsapparaten Till fullständigande af de meddelanden vi förut haft angående denna märkliga uppfin - ning lämna vi här en redogörelse för den er - farenhet man vunnit beträffande telegrafe - ring utan tråd och om den tillämpning den - samma kan erhålla vid styrning af torpeder Hrr Orlings ocli Braunerlijelms föregångare Den förste som framlade en teori att det ej består någon principiell skillnad mellan ljusets strålar och de från Leydenerfiaskan vid dennas urladdning utgående elektriska strå - larna var engelsmannen Maxwell Däref - ter påvisade tysken Hertz genom experiment att från elektriska oscillerande gnistor utbre - da sig svängningar eller vågor hvilka följa alldeles samma lagar som ljusvågorna Itali - enaren Calzecchi fann att rör med filspån uti erbjuda mycket olika motstånd mot den elektriska strömmen Fransmannen Branly

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