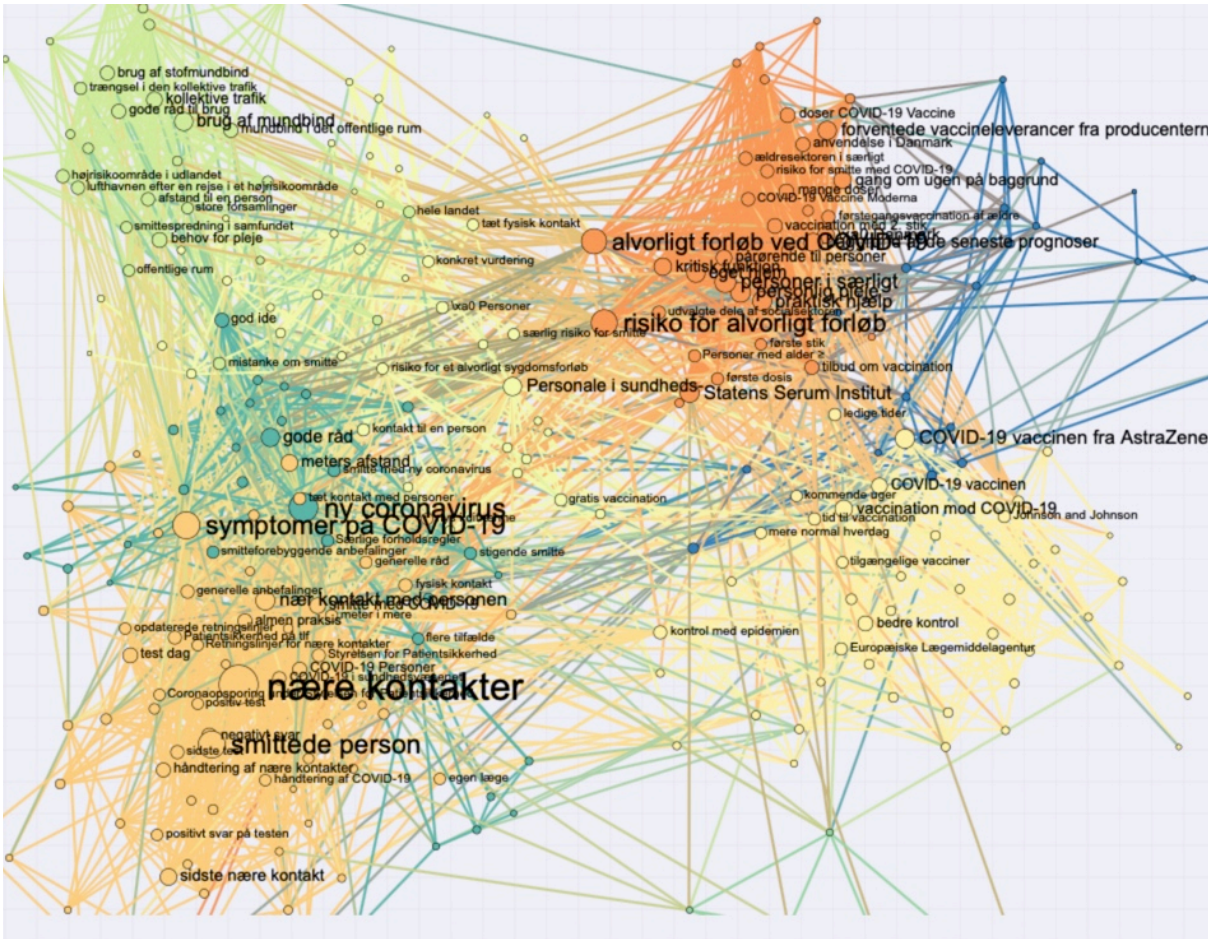
Sundhedsstyrelsens magt- og myndighedsudøvelse under Corona-epidemien gennem sproghandlinger i online nyheder, korreleret med data fra Statens Serum Institut

Forfatter: Thomas Tim Jensen
Charlottenlund, maj 2022



All rights reserved. Copyright© of Thomas Tim Jensen

Forord 3
Vision og mål for projektet 3
Leverede resultater 3
Opstilling af protokol - undersøgelsesdesign 3
Ønskede metoder og produkter 4
Indsamling af data - tabulære og og non-tabulære 4
Statsministerens pressemøder 4
Tallene fra SSI 4
Nyhedsteksterne fra SST 5
Særligt om scraping 5
Om HTML-tags 5
Om valg og oprensning af data - klar til datasprint 5
Sprint og produkter 6
Cortext 6
Netværksdiagram i Cortext 7
Pandemiens fase 1 og fase 2 8
Sammenhæng mellem semantisk-performativt indhold og HTML-tags 10
Tableau: Sammenhæng mellem testaktivitet og brug af HTML-tag <li> 11
BILAG: Script for scraping af SST-nyheder 12

All rights reserved. Copyright© of Thomas Tim Jensen

Forord

Denne rapport bygger på forskning, som jeg gennemførte i foråret 2022 på andet semester af masteruddannelsen i datadrevet organisationsudvikling (MDO) på Aalborg Universitet. Rapporten giver et indblik i mit undersøgelsesdesign og i de filosofiske forudsætninger, jeg har lagt ind i designet, idet jeg henviser til min baggrund som mag.art. i filosofi fra Københavns Universitet. Det bringer en række resultater bestående af såvel analyser som visualiseringer.

Der gælder almindelig ophavsret for tekst og visualiseringer, og der må ikke citeres fra teksten eller viderebringes visualiseringer uden henvisning til forfatter Thomas Tim Jensen, og denne rapport må ikke bringes i sin helhed uden forfatterens tilladelse.

Vision og mål for projektet

Hele verden gennemlevede i 2020 og 2021 Corona-/Covid-19-pandemien, en pandemi af et omfang og en karakter, som vel kun ganske få havde forudset eller vidste, hvordan de skulle håndtere undervejs.

Corona-pandemien som undersøgelsesgenstand har den store fordel set i perspektivet af AI og machine learning, at den er beskrevet med rigelige data af de datatyper, som jeg har kunnet træne mit dataprojekt på. Hos Sundhedsstyrelsen (fremover forkortet "SST") finder jeg non-tabulær, men html-opmærket tekst, som kan indgå i semantiske analyser, og hos Statens Serum Institut (fremover forkortet "SSI") findes tabulære data, begge dele opmærket med datostempler.

Leverede resultater

Mine resultater falder i to halvdele:

- For det første semantisk analyse og visualisering af nyhedsteksterne fra SST og
- for det andet visualisering af tal fra SSI "mashed" med nyhedsteksterne og med oversigt over testaktivitet sat i forhold til brug af liste-tag (<li> m.m. i nyhedernes HTML-kode).

Opstilling af protokol - undersøgelsesdesign

Mit undersøgelsesdesign tager udgangpunkt i to filosofiske antagelser:

1. For det første danner Aristoteles (384-322 f.Kr.) udgangspunkt for, hvad man kunne kalde protokollens hermeneutiske ramme i tid og rum. Jeg definerer pandemien som et drama ud fra det velkendte mønster fra Aristoteles' "Poetikken":
    1. en begyndelse: Statsminister Mette Frederiksens pressemøde i Statsministeriet den 11. marts 2020, hvor hun så at sige "indstifter" pandemien i Danmark (jf. indstiftelsesordene ved nadveren i kirken)
    2. en midte (klimaks/"turning point"): Vaccinationer går i gang den 27. december 2020
    3. en slutning: Statsminister Mette Frederiksens pressemøde i Statsministeriet den 26. januar 2022 (jf. udgangsbønnen i kirken)
    4. Jeg ønskede at opspore dette mønster i pandemiens forløb ved nærmere at betragte de to faser, altså fase 1 op til klimaks og fase 2 fra klimaks frem til pandemiens slutning.
2. For det andet antager jeg en sprogfilosofisk pointe fra den britiske filosof J.L. Austins (1911-1969) hovedværk "How To Do Things With Words"(1962), nemlig om sproghandlinger ("performatives"). SST's magt- og myndighedsudøvelse skete ikke mindst i deres nyhedsformidling på sst.dk, hvor de giver anbefalinger og påbud - som jo er sproghandlinger *par excellence*.
    1. I mine forundersøgelser til protokollen stod det mig tydeligt klart som en hensigtsmæssig tese, at SST's påbud og anbefalinger ofte blev givet i form af punkt- og nummerede lister - hvilket afspejlede sig i de online nyhedsteksters HTML-kode i form af opmærkning med -tags. Det var en tese, jeg ønskede at

All rights reserved. Copyright© of Thomas Tim Jensen

forfølge nærmere. Den arketypiske, bibelske punktliste er jo stentavlerne med de 10 bud.
2. Jeg ønskede også at se, om der var en sammenhæng mellem sproghandlinger og -tags og Corona-pandemiens dramatiske forløb.

3. For det tredje kan jeg tilføje endnu en filosofisk antagelse i mit undersøgelsesdesign, nemlig at SST's myndighedsudøvelse var tæt korreleret til pandemiens faktiske forløb, dokumenteret i tallene fra SSI, og altså at SST's myndighedsudøvelse skete i en vekselvirkning med virkeligheden. Jeg antager altså ikke en konstruktionisme, hvor SST's nyhedsformidling alene skulle kunne skabe/konstruere virkeligheden. Folk blev virkeligt syge, og nogle afgik ved døden af sygdommen. Dette kunne man kalde en form for filosofisk realisme.

Ønskede metoder og produkter

Med ovenstående antagelser som udgangspunkt overvejede jeg, hvilke metoder der bedst kunne indgå i mit undersøgelsesdesign, og hvilke produkter jeg kunne skabe derudaf. Her stod det mig klart, at jeg skulle se i retning af semantisk analyse og tilhørende visualisering for SST's nyhedsteksters vedkommende og i retning af Excel-forankret opstilling og analyse og Tableau-baseret visualisering, hvad angik tallene fra SSI.

Indsamling af data - tabulære og og non-tabulære

Jeg afgrænsede hurtigt mine datakilder til tre:

- Udskrifterne fra det indledende og det afsluttende pressemøde i Statsministeriet
- Tallene fra SSI var lige til at hente fra deres hjemmeside i let tilgængeligt format i Excel/CSV.
- Nyhedsteksterne fra SST måtte jeg derimod arbejde temmelig meget for at hente ned.

Statsministerens pressemøder

Den hermeneutiske ramme for dette dataprojekt er som sagt statsministerens indledende og afsluttende pressemøder.

- **Indledende pressemøde**. En gennemlæsning af transskriptionen fra pressemødet fik mig til at notere følgende nøgleord: skræmmebillede, smitte, epidemi, fællesskab, kapacitet, sundhedsfaglig, sårbare, samfundssind, myndigheder, lukke, forhindre, spørgsmål, svar. https://www.stm.dk/presse/pressemoedearkiv/pressemoede-om-covid-19-den-11-marts-2020/
- **Afsluttende pressemøde**. Jeg noterede mig her følgende ord: milepæl, overgang, smittetal, fællesskab, vaccine, hensyn, varianter, sundhedsvæsen, forår. https://www.stm.dk/presse/pressemoedearkiv/pressemoede-onsdag-den-26-januar-2022/

Andre ville måske have lagt mærke til andre ord. Såfremt tiden og pladsen havde tilladt det, og hvis det havde været mere relevant for dette dataprojekt, kunne man have valgt at gå dybere litteraturvidenskabeligt til værks og have udsat de to transskriptioner for en litterær analyse, og en såkaldt aktant-analyse kunne her have været relevant. Her og nu nøjes jeg med intuitivt at samle og anvende ovenstående nøgleord til at give en pejling for mit dataprojekt, ikke mindst for den "story", som jeg ønsker, at mine data og visualiseringer skal fortælle.

Tallene fra SSI

Det meget omfattende talmateriale fra SSI var heldigvis meget let tilgængeligt og i gængse filformater, som afspejlede den begrebs- og datamodel, som SSI ret hurtigt havde på plads fra pandemiens begyndelse: https://covid19.ssi.dk/overvagningsdata/vaccinationstilslutning

All rights reserved. Copyright© of Thomas Tim Jensen

Nyhedsteksterne fra SST

Straks sværere var det med nyhedsteksterne. Hvordan skulle jeg "scrape" nyhederne for såvel tekstindhold som deres tekstrelaterede HTM. Der opstod endvidere den vanskelighed, at oversigtssiden for nyheder på SST's hjemmeside omkring den tid, da jeg skulle bruge dataene, var utilgængelig (https://sst.dk/da/Nyheder). Jeg valgte da straks at kontakte Droids Agency for at bede om assistance til at skrive et egnet Python-script. Dette skete på et tre timers møde mandag den 11. april 2022.

Særligt om scraping

Jeg kontaktede som sagt bureauet Droids Agency, og en af deres dygtige data scientists sparrede og rådgav mig i forhold til at "scrape" de cirka 400 nyheder fra SST's hjemmeside.

Jeg havde forud defineret det feature space, som jeg fandt hensigtsmæssigt for mit undersøgelsesdesign, særligt datoopmærkning, tekstindhold og HTML-opmærkning af nyhederne, og konsulenten fra Droids Agency skrev i en Jupyter-notebook det Python-script, som gjorde det muligt at scrape nyhederne og lægge dem ned i en Excel-fil. Scriptet ligger som dokumentation i bilag.

Udfordringen bestod dels i, at oversigtssiden for nyheder på sst.dk i dagene, hvor jeg skulle scrape, som sagt ikke var tilgængelig. Konsulenten kunne med kreativ brug af sin erfaring foretage en slags "reverse engineering" af, hvorfra nyhederne skulle hentes, herunder SST's API, dels at konsulenten og jeg sammen måtte gennemgå kildekoden til nyhederne for at udpege de opmærkninger, som var relevante - jeg havde identificeret de væsentligste og tilsendt konsulenten dem forud for mødet.

Jeg efterbehandlede og gennemgik manuelt de scrapede data for at danne mig et overblik som forberedelse til datasprint på AAU den 20. - 22. april 2022. Det viste sig dér, at mine data var gode og næsten tilstrækkeligt præparerede til, at jeg kunne bygge min pipeline - eneste vanskelige punkt var at sammenkøre datoopmærkede data fra hhv. SSI og SST i én tabel - dette problem løste jeg dagen efter datasprintet, så problemet var ikke så stort, som jeg havde frygtet.

Om HTML-tags

Undervejs var det vigtigt for konsulenten fra Droids Agency og mig at finde et godt niveau for granulariteten for scrapingen - ikke for mange kolonner, men heller ikke for få. Det var interessant for mig efter scrapingen at se nyhedsteksterne opstillet i tabelformat, altså i kolonner og rækker, og altså se en form for røntgenfoto af nyhedsteksterne, hvorudfra jeg intuitivt kunne stille en form for diagnose eller karakteristik af teksterne som fikspunkter for min efterfølgende maskinelle analyse.
Liste-tags'ene kom i forskellige former: ordered_lists, lists, factbox text, factbox list. Efter en manuel gennemgang af alle nyhederne besluttede jeg at definere brug af <li> til kun at dække nogle af disse former - og så fastholdt jeg denne afgrænsning formalistisk, således at hvis der var indhold opmærket med disse tags, så var indholdet pr. definition opmærket med -tag, og det var velsagtens semantisk korrekt for 95 % vedkommende, samtidig med at jeg kunne fastholde en stringens og en objektivitet i anvendelsen af min definition.

Om valg og oprensning af data - klar til datasprint

Jeg var åbent indstillet over for, hvordan datasprinten måtte forløbe, og hvad den måtte kaste af sig, så jeg samlede mine data sammen, så godt det kunne lade sig gøre, og uden at behandle dem for meget manuelt, såfremt der på sprinten skulle vise sig en automatiseret måde at gøre noget, som ville tage mig halve dage manuelt.

All rights reserved. Copyright© of Thomas Tim Jensen

Sprint og produkter

Sprinten blev en stor succes for mig, takket være et overskueligt datasæt, et forholdsvis velafgrænset og udvalgt antal metoder og værktøjer, og to forholdsvis spidsvinklede forskningsspørgsmål:

- "Hvad kan sammenhængen mellem HTML-liste-tags og det semantiske indhold i SST's anbefalinger sige om SST's sprogformidlede myndighedsudøvelse, korreleret til SSI's data for epidemien?" og
- "Afspejles pandemiens dramatiske forløb i to faser i en tilsvarende network mapping?"

Det eneste, jeg ikke nåede på datasprinten var at korrelere hyppigheden af liste-tags med testaktiviteten. Dette arbejdede jeg mig dog frem til dagen efter datasprinten.

Jeg vil nu vise og annotere de visualiseringer, som jeg har dannet mig. De fordeler sig på henholdsvis arbejde udført i Cortext og i Tableau.

Cortext

Min arbejdsgang i Cortext var følgende:

1. Lægge al tekst fra mine høstede nyheder fra sst.dk i Excel-ark
2. Gemme som en csv-fil - zippe
3. Oprette projekt i Cortext
4. Uploade zip-fil til Cortext
5. Gå i gang med projektet:
   1. Data parsing
   2. Terms extraction
   3. Network mapping
   4. Heat mapping

Jeg ønskede som sagt at undersøge, hvilke temaer SST's nyhedstekster omhandlede, og om der kunne ses et tydeligt skifte fra første til anden fase af pandemien - med igangsættelsen af vaccinerne som vendepunkt. Jeg ønskede også at undersøge sammenhængen mellem temaerne og HTML-opmærkningen.

Derefter ville jeg, i Tableau, se, hvilken sammenhæng der kunne være mellem HTML-listeopmærkningen og pandemiens forløb. På de næste sider vil jeg præsentere og kommentere de visualiseringer, jeg skabte på datasprinten.

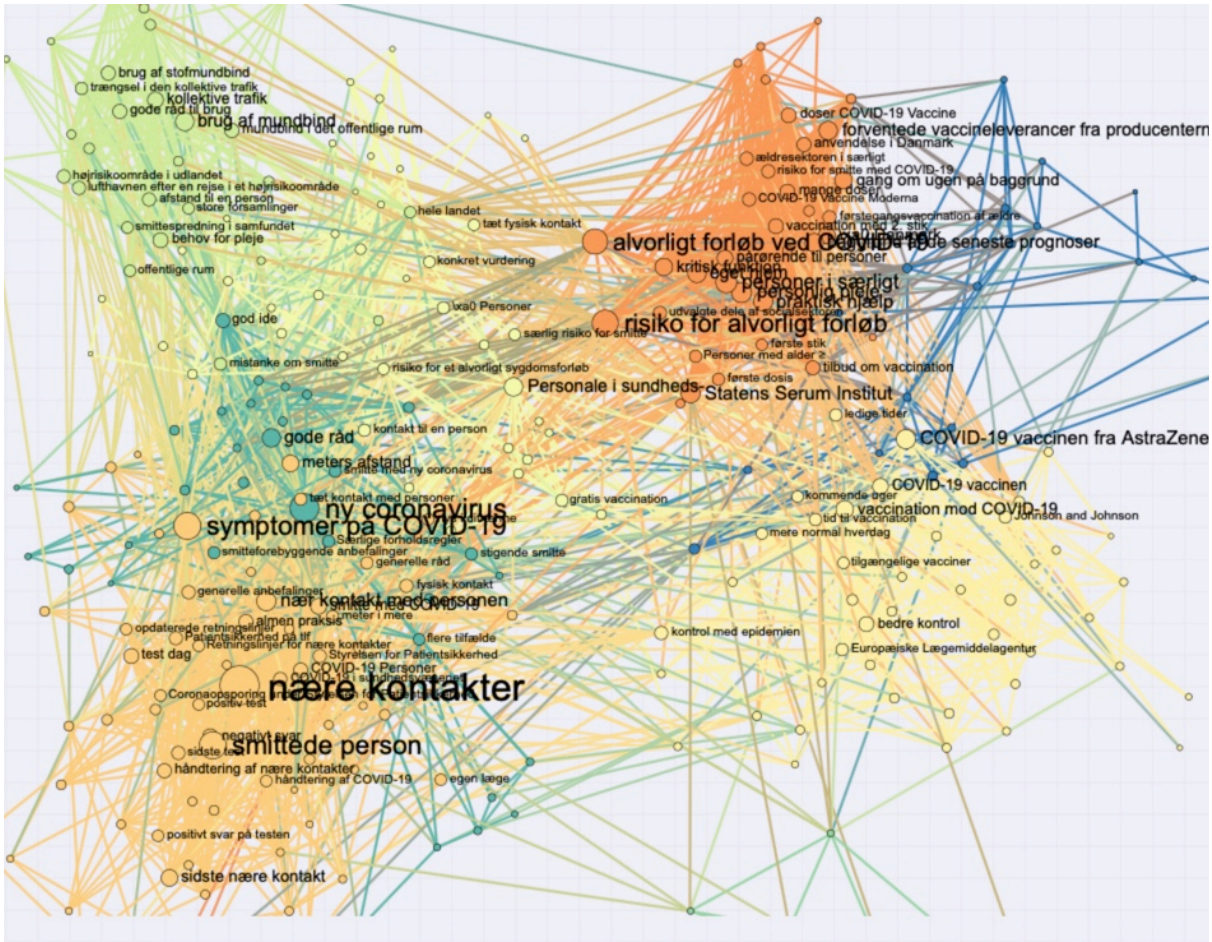
All rights reserved. Copyright© of Thomas Tim Jensen

Netværksdiagram i Cortext

Netværksdiagrammet i Cortext over alle SST's nyheder i pandemien viser forholdsvis klare "klynger"/"clusters" i termerne og med mange edges mellem noderne. Temaerne fordeler sig således:

- Mundbind og færdsel i det offentlige rum (lysegrøn)
- Adfærd i det personlige rum (mørkegul)
- Gode råd (grøn) - forbinder lysegrøn og mørkegul
- Om sygdommen (orange)
- Om vaccinen (gul)

Særligt lægger jeg mærke til, at klyngen "Gode råd" forbinder emnerne om ønsket adfærd i henholdsvis det offentlige og det private, nære rum. Dette indikerer, at mit undersøgelsesdesign har været hensigtsmæssigt opstillet.



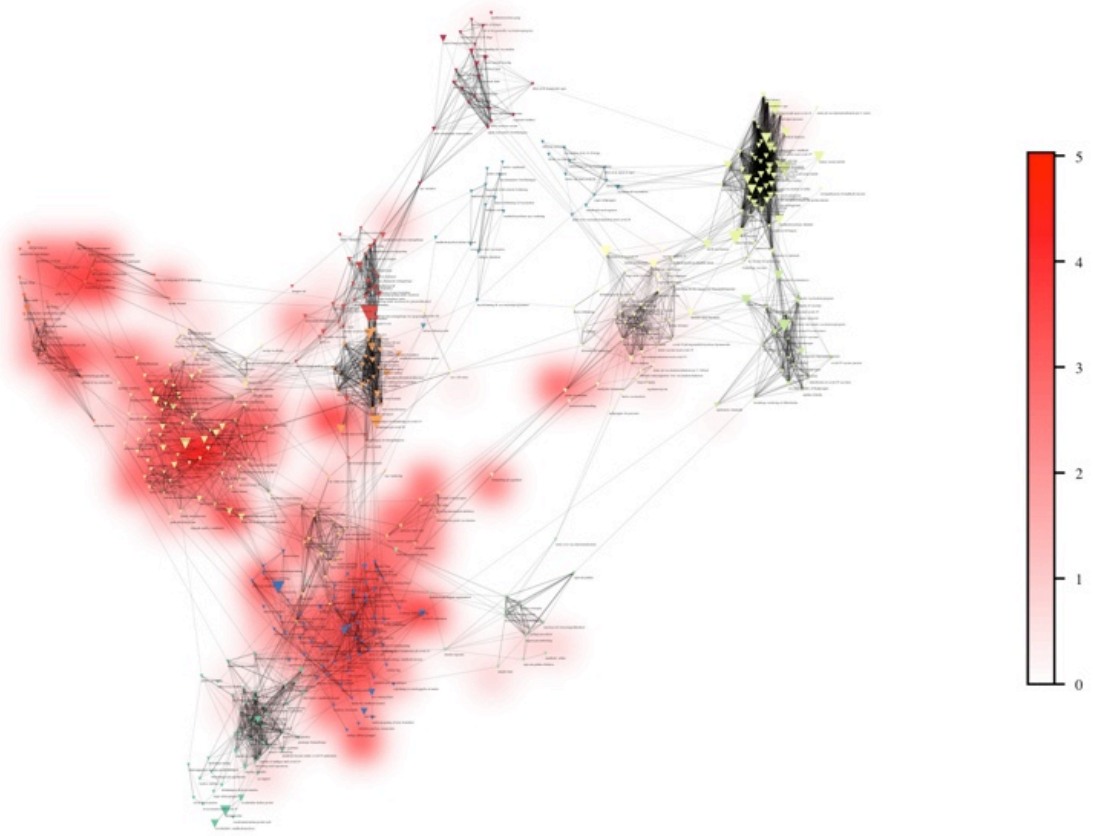
All rights reserved. Copyright© of Thomas Tim Jensen

Pandemiens fase 1 og fase 2

Her kommer vi til den del af undersøgelsen, som var stærkt inspireret af Aristoteles. Jeg delte mine nyheder op i to halvdele: Den første halvdel med nyheder fra pandemiens begyndelse frem til vendepunktet (= fase 1), den anden halvdel med nyheder fra vendepunktet og frem til pandemiens afslutning (= fase 2). Netværksdiagrammet fra før blev mit "grundkort" over pandemiens geografi, metaforisk forstået, og oven på dette lagde jeg så først en kalke med et "heatmap" over fase 1, dernæst fase 2.

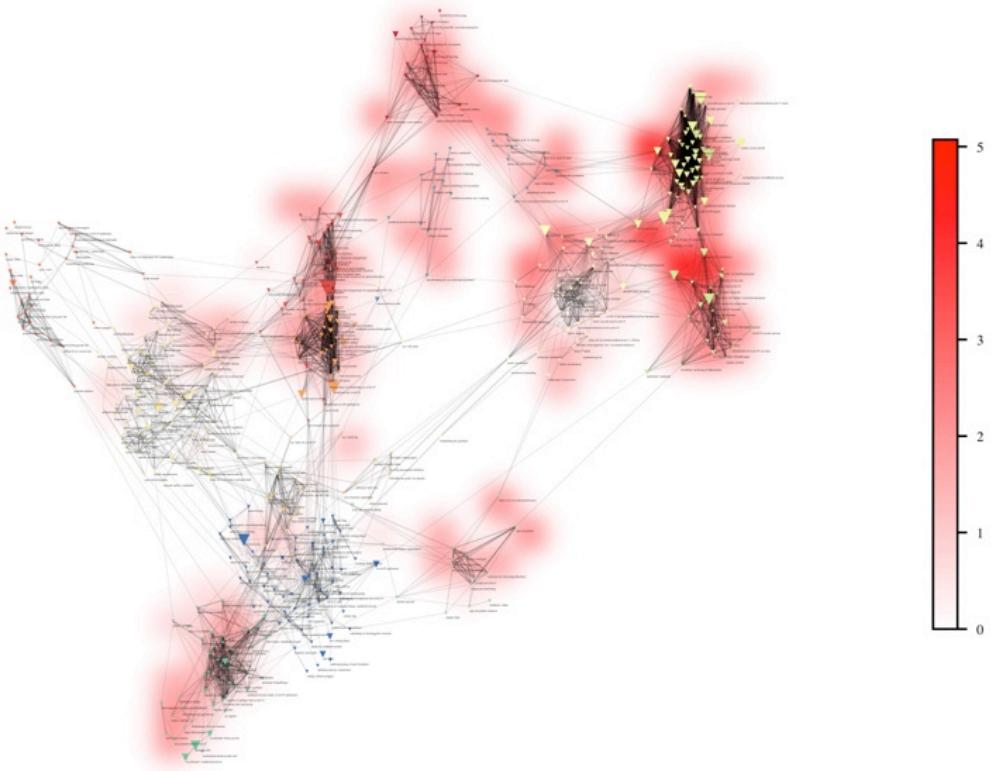
Da disse heatmaps trådte frem, var det en meget positiv overraskelse, at faserne adskilte sig så tydeligt. Det indikerede, at mit undersøgelsesdesign også for denne del var hensigtsmæssig. Med et glimt i øjet kunne jeg (som mag.art. i filosofi) endnu engang konstatere, at "Aristoteles har altid ret"!

- I pandemiens **fase 1** træder disse emner tydeligt frem: mundbind og færdsel i det offentlige rum, adfærd i det personlige rum, gode råd:



All rights reserved. Copyright© of Thomas Tim Jensen

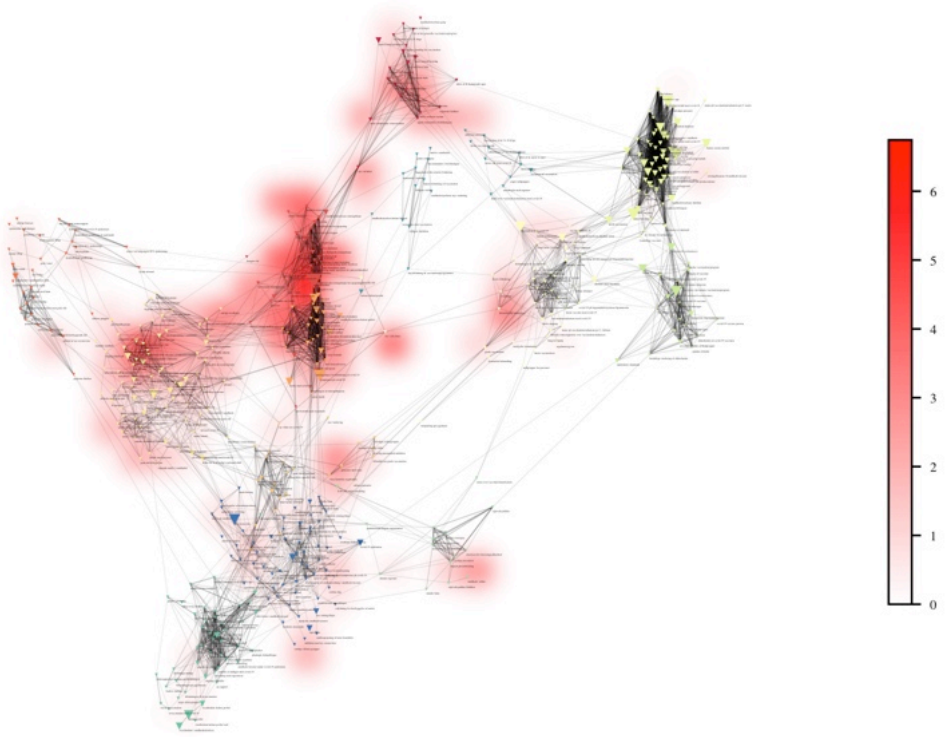
- I pandemiens **fase 2** træder derimod disse emner tydeligt frem: om sygdommen og om vaccinen.



All rights reserved. Copyright© of Thomas Tim Jensen

Sammenhæng mellem semantisk-performativt indhold og HTML-tags

Afslutningsvis i Cortext, og med samme "grundkort", dannede jeg et heatmap over forekomsten af HTML-listetags i forbindelse med temaerne i samtlige SST's nyheder. Her blev resultatet også, at mit undersøgelsesdesign forekom hensigtsmæssigt, idet heatmappet viste, at listetagget særligt blev brugt i forbindelse med temaerne mundbind og færdsel i det offentlige rum og adfærd i det personlige rum samt gode råd herom - plus yderligere i forbindelse med råd til institutioner med videre. Quod erat demonstrandum.



All rights reserved. Copyright© of Thomas Tim Jensen

Tableau: Sammenhæng mellem testaktivitet og brug af HTML-tag

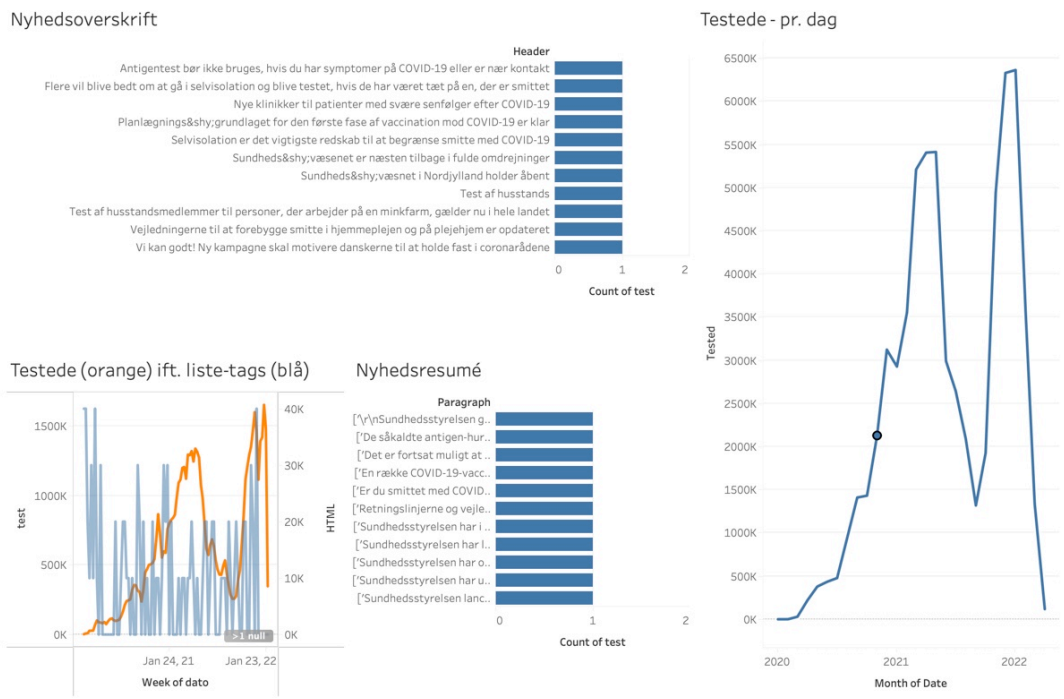
Endelig byggede jeg på datasprinten et dashboard i Tableau, hvor jeg ønskede at vise et kurvediagram over testaktiviteten, SST's nyheder og brugen af liste-tag.
Grunden til, at jeg fandt testaktiviteten interessantest i denne fremstilling af SST's magt- og myndighedsudøvelse under pandemien, var, at testaktiviteten var en borgeradfærd, som SST direkte kunne påvirke, hvor smittetryk og sygdomstilfælde kun kunne siges at være afledede og indirekte påvirkelige fænomener, som nyhedsformidling ikke kan tale op eller ned.

Jeg importerede min CSV-fil med SST-nyheder fra min session med Droids Agency samt en CSV-fil med de relevante tal fra SSI.

I dashboardet nedenfor ser man testaktiviteten i kurvediagrammet til højre. Det er klikbart, og dér, hvor man klikker på kurven, får man i de to vinduer med vandrette streger vist de tilhørende nyheder fra samme måned, øverst som overskrift, nedenunder som nyhedsresumé.

I det sidste vindue vises (med den orange kurve) samme testaktivitet, men nu med et lag (med blåt) over den større eller mindre brug af liste-tagget i SST's nyheder. Jeg har indrettet diagrammet med to værdiakser - og jeg har ganget forekomsten af liste-tags med faktor 10.000, således at forekomsten træder tydeligere frem.
Det var i øvrigt ikke helt ligetil at få lagt SST-nyhederne på samme tidskolonne som SSI's tal, men med en del sortering og manuel oprydning lykkedes det til sidst.

Hvis jeg skulle tolke diagrammet, kunne jeg forsigtigt foreslå at tolke det som, at SST i pandemiens begyndelse, hvor testaktiviteten var lav, ret hyppigt brugte liste-tags, og dermed - hvilket vi ved fra heatmappene i Cortext ovenfor - ønskede at udøve myndighed temmelig kraftigt, og i takt med at testaktiviteten steg kraftigt, faldt brugen af liste-tags om ikke tilsvarende, så dog noget, indtil den steg igen i pandemiens anden fase, da man ønskede at påvirke befolkningen til at lade sig vaccinere.



All rights reserved. Copyright© of Thomas Tim Jensen

BILAG: Script for scraping af SST-nyheder


```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 94,
   "id": "d26ae626",
   "metadata": {},
   "outputs": [],
   "source": [
    "import requests\n",
    "from bs4 import BeautifulSoup\n",
    "import pandas as pd\n",
    "import json\n",
    "#https://www.sst.dk/api/search/Search?
draw=1&columns%5B0%5D%5Bdata%5D=function&columns%5B0%5D%5Bname%5D=
&columns%5B0%5D%5Bsearchable%5D=true&columns%5B0%5D%5Borderable%5D=fa
lse&columns%5B0%5D%5Bsearch%5D%5Bvalue%5D=&columns%5B0%5D%5Bsearch
%5D%5Bregex%5D=false&start=0&length=100&search%5Bvalue%5D=&search%5Bregex
%5D=false&filters=%7BAF884A43-9989-44B1-8A29-352E72D3325A%7D&searchWord=&c
ontextId=%7B07B7F301-0A10-49D5-
A8A8-43EBF0D4EE28%7D&searchType=news&pageSize=100&_=1649425656780\n",
    "pay = {\"draw\":1,\n",
    "    \"columns%5B0%5D%5Bdata%5D\":\"function\",\n",
    "    \"columns%5B0%5D%5Bname%5D\":\"\",\n",
    "    \"columns%5B0%5D%5Bsearchable%5D\":\"true\",\n",
    "    \"columns%5B0%5D%5Borderable%5D\":\"false\",\n",
    "    \"columns%5B0%5D%5Bsearch%5D%5Bvalue%5D\":\"\",\n",
    "    \"columns%5B0%5D%5Bsearch%5D%5Bregex%5D\":\"false\",\n",
    "    \"start\":0,\n",
    "    \"length\":100,\n",
    "    \"search%5Bvalue%5D\":\"\",\n",
    "    \"search%5Bregex%5D\":\"false\",\n",
    "    \"filters\":\"%7BAF884A43-9989-44B1-8A29-352E72D3325A%7D\",\n",
    "    \"searchWord\":\"\",\n",
    "    \"contextId\":\"%7B07B7F301-0A10-49D5-A8A8-43EBF0D4EE28%7D\",\n",
    "    \"searchType\":\"news\",\n",
    "    \"pageSize\":100}\n",
    "draw = 1\n",
    "start = 0\n",
    "pageSize = 100\n",
    "link = \"https://www.sst.dk/api/search/Search?\"\n",
    "base = \"https://www.sst.dk\"\n",
    "frame = []\n",
    "while True:\n",
    "    pay['draw'] = draw\n",
    "    pay['start'] = start\n",
    "    search = link + \"&\".join([\"=\".join([key, str(item)]) for key, item in pay.items()])\n",
    "    r = requests.get(search)\n",
    "    data = json.loads(r.text)['data']\n",
    "    if len(data)<=1:break\n",
    "    frame += data\n",
    "    draw += 1\n",
    "    start += pageSize\n",
    "\n",
```


All rights reserved. Copyright© of Thomas Tim Jensen

```
    "links = pd.DataFrame(frame, columns=['overskift', 'dato', 'header', 'link', 'none',
'image'])\n",
    "links['link'] = links.link.apply(lambda x: \"\".join([base, x]))"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 113,
   "id": "164c3897",
   "metadata": {},
   "outputs": [
    {
     "data": {
      "text/plain": [
       "['De gode råd til at forebygge coronasmitte virker også mod influenza',\n",
       " '25 MAR 2022',\n",
       " 'Sundhedsstyrelsens seneste statusrapport viser, at der i øjeblikket er en sen
influenzasæson i Danmark. Mens antallet af personer, der tester positiv for covid-19, er
faldet, stiger antallet af personer med influenza. Andre europæiske lande har også en sen
influenzasæson. ',\n",
       " 'https://www.sst.dk/da/Nyheder/2022/De-gode-raad-til-at-forebygge-coronasmitte-
virker-ogsaa-mod-influenza',\n",
       " None,\n",
       " 'o-search-result-list__image-container--no-image']"
      ]
     },
     "execution_count": 113,
     "metadata": {},
     "output_type": "execute_result"
    }
   ],
   "source": []
  },
  {
   "cell_type": "code",
   "execution_count": 100,
   "id": "9e9387c5",
   "metadata": {},
   "outputs": [],
   "source": [
    "saved = links.copy()\n",
    "saved.to_csv('links.csv')"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 132,
   "id": "021af36e",
   "metadata": {
    "scrolled": true
   },
   "outputs": [
    {
     "name": "stdout",
     "output_type": "stream",
     "text": [
      "0 / 409\n",
      "10 / 409\n",
```


All rights reserved. Copyright© of Thomas Tim Jensen

```
    "20 / 409\n",
    "30 / 409\n",
    "40 / 409\n",
    "50 / 409\n",
    "60 / 409\n",
    "70 / 409\n",
    "80 / 409\n",
    "90 / 409\n",
    "100 / 409\n",
    "110 / 409\n",
    "120 / 409\n",
    "130 / 409\n",
    "140 / 409\n",
    "150 / 409\n",
    "160 / 409\n",
    "170 / 409\n",
    "180 / 409\n",
    "190 / 409\n",
    "200 / 409\n",
    "210 / 409\n",
    "220 / 409\n",
    "230 / 409\n",
    "240 / 409\n",
    "250 / 409\n",
    "260 / 409\n",
    "270 / 409\n",
    "280 / 409\n",
    "290 / 409\n",
    "300 / 409\n",
    "310 / 409\n",
    "320 / 409\n",
    "330 / 409\n",
    "340 / 409\n",
    "350 / 409\n",
    "360 / 409\n",
    "370 / 409\n",
    "380 / 409\n",
    "390 / 409\n",
    "400 / 409\n"
   ]
  }
 ],
 "source": [
  "raw_html = []\n",
  "for k, corona_link in enumerate(saved.itertuples()):\n",
  "    if k % 10 == 0:\n",
  "        print(k, '/', len(saved))\n",
  "    r = requests.get(corona_link.link)\n",
  "    if r.status_code == 200:\n",
  "        raw_html.append(r.text)\n",
  "    else:\n",
  "        raw_html.append('error')"
 ]
},
{
 "cell_type": "code",
 "execution_count": 136,
 "id": "c8abb9cf",
```


All rights reserved. Copyright© of Thomas Tim Jensen

```
  "metadata": {},
  "outputs": [],
  "source": [
   "saved['html'] = raw_html"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 137,
  "id": "990a9db2",
  "metadata": {},
  "outputs": [],
  "source": [
   "saved.to_csv('html_raw.csv')"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 143,
  "id": "4fe587c7",
  "metadata": {},
  "outputs": [],
  "source": [
   "soup = BeautifulSoup(saved.html[0], 'html.parser')"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 150,
  "id": "909a265e",
  "metadata": {},
  "outputs": [
   {
    "data": {
     "text/plain": [
      "'https://www.sst.dk/da/Nyheder/2022/De-gode-raad-til-at-forebygge-coronasmitte-
virker-ogsaa-mod-influenza'"
     ]
    },
    "execution_count": 150,
    "metadata": {},
    "output_type": "execute_result"
   }
  ],
  "source": [
   "saved.link[0]"
  ]
 },
 {
  "cell_type": "code",
  "execution_count": 218,
  "id": "8f12e4b1",
  "metadata": {
   "scrolled": true
  },
  "outputs": [
   {
    "name": "stdout",
```


All rights reserved. Copyright© of Thomas Tim Jensen

```
  "output_type": "stream",
  "text": [
   "0 / 409\n",
   "10 / 409\n",
   "20 / 409\n",
   "30 / 409\n",
   "40 / 409\n",
   "50 / 409\n",
   "60 / 409\n",
   "70 / 409\n",
   "80 / 409\n",
   "90 / 409\n",
   "100 / 409\n",
   "110 / 409\n",
   "120 / 409\n",
   "130 / 409\n",
   "140 / 409\n",
   "150 / 409\n",
   "160 / 409\n",
   "170 / 409\n",
   "180 / 409\n",
   "190 / 409\n",
   "200 / 409\n",
   "210 / 409\n",
   "220 / 409\n",
   "230 / 409\n",
   "240 / 409\n",
   "250 / 409\n",
   "260 / 409\n",
   "270 / 409\n",
   "280 / 409\n",
   "290 / 409\n",
   "300 / 409\n",
   "310 / 409\n",
   "320 / 409\n",
   "330 / 409\n",
   "340 / 409\n",
   "350 / 409\n",
   "360 / 409\n",
   "370 / 409\n",
   "380 / 409\n",
   "390 / 409\n",
   "400 / 409\n"
  ]
 }
],
"source": [
 "tagged_information = []\n",
 "for k, page in enumerate(saved.itertuples()):\n",
 "    if k % 10 == 0:\n",
 "        print(k, '/', len(saved))\n",
 "    soup = BeautifulSoup(page.html, 'html.parser')\n",
 "\n",
 "    # overskrift\n",
 "    h1 = [a.text for a in soup.find_all('h1')]\n",
 "\n",
 "    # header\n",
 "    h2 = [a.text for a in soup.find_all('h2')]\n",
```


All rights reserved. Copyright© of Thomas Tim Jensen

"\n",
"   # paragraph\n",
"   banned = ['Nyheder, udgivelser og arrangementet fremgår desværre ikke via menuerne. Brug søgefunktionen i stedet: Søg']\n",
"   html_p = [a for a in soup.find_all('p') if a.text not in banned][:-3]\n",
"   paragraph_text = [a.text for a in html_p]\n",
"\n",
"   # links\n",
"   html_a = [a for paragraph in html_p for a in paragraph.find_all('a')]\n",
"   link_text = [a.text for a in html_a]\n",
"\n",
"   # clean up paragraph text\n",
"   paragraph_text = [a for a in paragraph_text if a not in link_text]\n",
"\n",
"   # ol\n",
"   ordered_lists = [f\"{k+1}. {b.text}\" for a in soup.find_all('ol') for k, b in enumerate(a.find_all('li'))]\n",
"\n",
"   # li\n",
"   lists = [a.text for b in soup.find_all('article') for a in b.find_all('li') if len(a.text.split()) > 1]\n",
"\n",
"   # factbox\n",
"   class_name = 'm-factbox'\n",
"   factbox_text = [a.text.replace('\\n', ' ').strip() for a in soup.find_all('div', {\"class\": class_name})]\n",
"\n",
"   # factbox list\n",
"   factbox_list = [b.text for a in soup.find_all('div', {'class': class_name}) for b in a.find_all('li')]\n",
"\n",
"\n",
"   tagged_text = {'h1': h1,\n",
"   'h2': h2,\n",
"   'paragraph': paragraph_text,\n",
"   'link_text': link_text,\n",
"   'ordered_lists': ordered_lists,\n",
"   'list': lists,\n",
"   'factbox_text': factbox_text,\n",
"   'factbox_list': factbox_list}\n",
"   \n",
"   tagged_information.append(tagged_text)"
]
},
{
"cell_type": "code",
"execution_count": 220,
"id": "4e0f3573",
"metadata": {},
"outputs": [
{
"data": {
"text/plain": [
"{'h1': ['Sundheds\\xadvæsnet i Nordjylland holder åbent'],\n",
" 'h2': ['Det er fortsat muligt at gå til lægen og få behandling på sygehuset, selvom du bor i en af de syv nordjyske kommuner, hvor der er lokale restriktioner og tiltag. ',\n",
" 'Sundhedspersonale kan fortsat tage på arbejde'],\n",

All rights reserved. Copyright© of Thomas Tim Jensen

" 'paragraph': ['For at undgå smittespredning i 7 nordjyske kommuner har regeringen besluttet at indføre en række lokale restriktioner og tiltag i de ramte kommuner. Midt i den alvorlige situation er det vigtigt for Sundhedsstyrelsen at understrege, at det fortsat er muligt at gå til sin praktiserende læge, hvis man er syg, at få behandling på sygehuset, og at sundhedspersonale stadig skal tage på arbejde.\\xa0',\n",
" 'Det er en anden situation nu end i foråret, hvor store dele af sundhedsvæsenet lukkede ned. Det skyldtes bekymring for, om der kunne opstå problemer med kapaciteten. Men i den nuværende situation er der ikke kapacitetsproblemer i Region Nordjylland, og samtidig har sundhedsvæsnet stort fokus på smitteforebyggelse.',\n",
" 'I udgangspunktet bliver alle ydelser i sundhedsvæsenet, bortset fra kosmetisk behandling, betragtet som kritiske funktioner. Patienter vil derfor også stadig kunne modtage behandling i almen praksis, i kommuner, hos tandlæger, fysioterapeuter, kiropraktorer, fodterapeuter mv.\\xa0',\n",
" '"Det er meget vigtigt, at befolkningen i de syv kommuner fortsat går til lægen, hvis de er syge, og de skal kunne få deres behandling på et sygehus, hvis det er nødvendigt. Det er udelukkende kosmetiske behandlinger, som skal udskydes, fx botox-behandling og tandblegning," siger vicedirektør Helene Probst.',\n",
" 'Patienter fra de 7 nordjyske kommuner kan også modtage behandling i de omkringliggende kommuner.',\n",
" 'Ansatte på sygehuse og plejehjem skal fortsat gå på arbejde, også selvom de bor i én kommune og arbejder i en anden. Det er afgørende, at vi holder sundhedsvæsnet åbent, forklarer Helene Probst:',\n",
" '"Langt de fleste ydelser i sundhedsvæsenet kan fortsætte. I mange tilfælde kan det nemlig have alvorlige konsekvenser for patienter og borgeres helbred og funktionsniveau, hvis en behandling bliver udskudt. Derfor er det naturligvis også vigtigt, at sundhedspersonale møder frem på arbejde – uanset hvor de bor eller arbejder. Samtidig er det vigtigt at sikre, at aktiviteter i sundheds- og ældresektoren sker under de rammer, rutiner og vaner for smitteforebyggelse, der er opbygget under COVID-19 epidemien. Man kan også fortsat benytte sig af fx videokonsultation mv.," siger Helene Probst.'],\n",
" 'link_text': [],\n",
" 'ordered_lists': [],\n",
" 'list': [],\n",
" 'factbox_text': [],\n",
" 'factbox_list': []}"
]
},
"execution_count": 220,
"metadata": {},
"output_type": "execute_result"
}
],
"source": [
"tagged_information[231]"
]
},
{
"cell_type": "code",
"execution_count": 225,
"id": "64c1a230",
"metadata": {},
"outputs": [],
"source": [
"tagged_frame = pd.DataFrame(tagged_information)\n",
"saved['h1'] = tagged_frame['h1']\n",
"saved['h2'] = tagged_frame['h2']\n",
"saved['paragraph'] = tagged_frame['paragraph']\n",
"saved['ordered_lists'] = tagged_frame['link_text']\n",
"saved['lists'] = tagged_frame['list']\n",

All rights reserved. Copyright© of Thomas Tim Jensen

```
    "saved['factbox_text'] = tagged_frame['factbox_text']\n",
    "saved['factbox_list'] = tagged_frame['factbox_list']"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 227,
   "id": "38104713",
   "metadata": {},
   "outputs": [],
   "source": [
    "saved.to_csv('formatted.csv')"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 228,
   "id": "dbf0523c",
   "metadata": {},
   "outputs": [],
   "source": [
    "saved.to_excel('formatted.xlsx')"
   ]
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 3 (ipykernel)",
   "language": "python",
   "name": "python3"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.9.7"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 5
}
```


***

All rights reserved. Copyright© of Thomas Tim Jensen