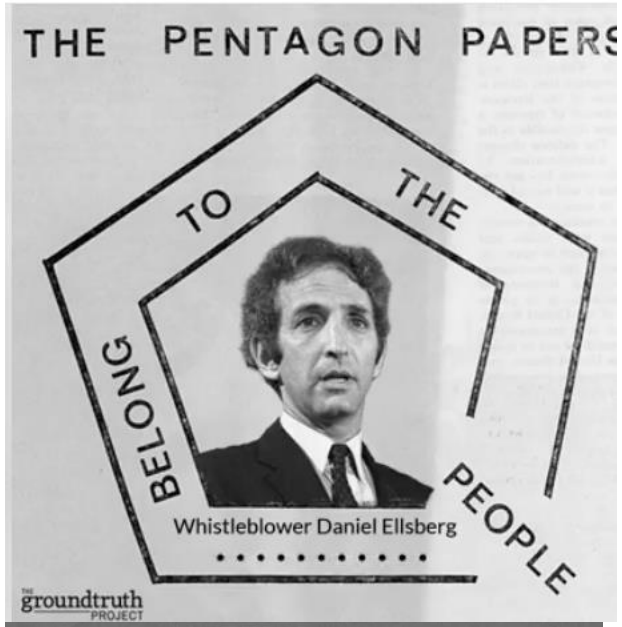
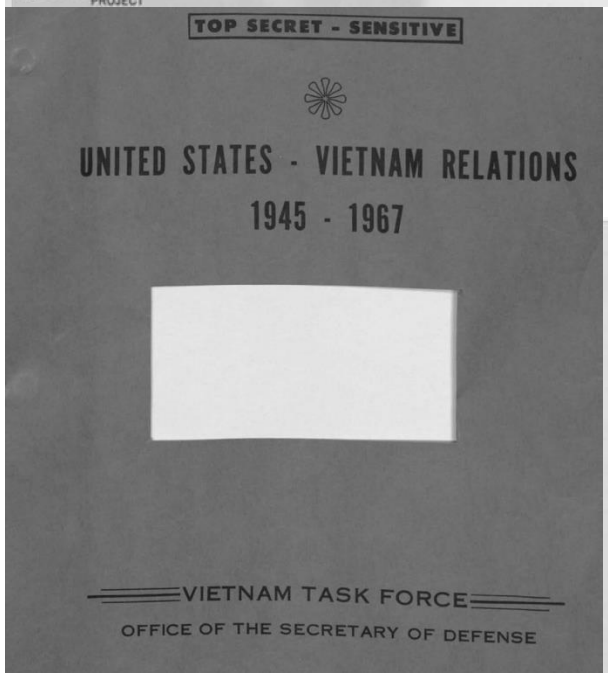
Projektet



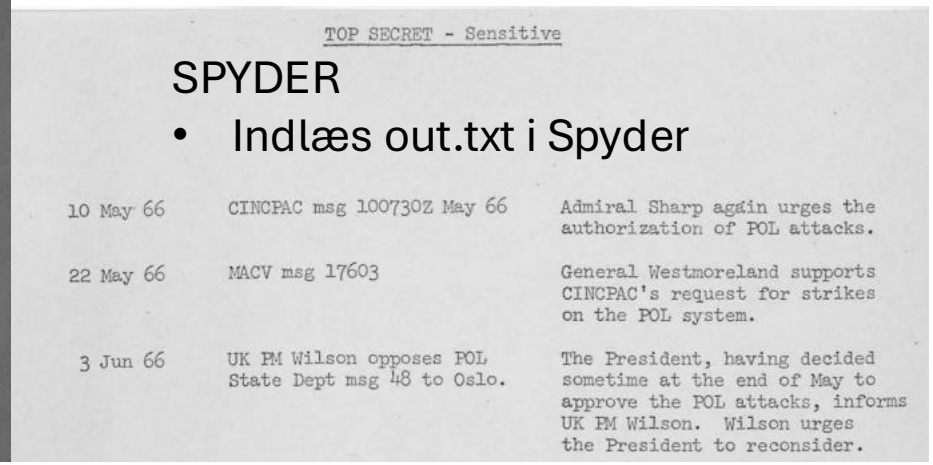


- https://www.archives.gov/research/pentagon-papers
- Download en pdf
  - cd Downloads
  - mkdir tmp
  - cd tmp
  - Flyt pdf'en til tmp-folderen

TERMINALEN

Installer poppler, imagemagic og tesseract

- pdfseparate Pentagon-Papers-Part-IV-C-7-b.pdf page-%d.pdf
- ls *pdf | while read x; do convert -density 300 $x $x.png; done
- ls *png | while read x; do tesseract $x $x; done
- LØS SORTERINGSPROBLEMET vha Python
- ls *txt | while read x; do cat $x >> out.txt; done


TOP SECRET - Sensitive

10 May 66	CINCPAC msg 100730Z May 66	Admiral Sharp again urges the authorization of POL attacks.
22 May 66	MACV msg 17603	General Westmoreland supports CINCPAC's request for strikes on the POL system.
3 Jun 66	UK PM Wilson opposes POL State Dept msg 48 to Oslo.	The President, having decided sometime at the end of May to approve the POL attacks, informs UK PM Wilson. Wilson urges the President to reconsider.

SPYDER

- Indlæs out.txt i Spyder

LØS SORTERINGSPROBLEMET vha Python

```
page-1.pdf
page-10.pdf
page-100.pdf
page-101.pdf
page-102.pdf
page-103.pdf
```

1. Lav en liste med filnavne i folderen med separerede *pdf-filer
2. Loop igennem listen og
    1. Identificér filer med kun ét tal og erstat med 00 (page-1 til page-001)
    2. Gør det samme for filer med to tal og erstat med 0 (page-10 til page-010)
3. Modificer trin 2 så du får en tekststreng pr filnavn på følgende form: ”mv page-1.pdf page-001.pdf”
4. Udskriv listen til en fil.

Projektet Organiseres

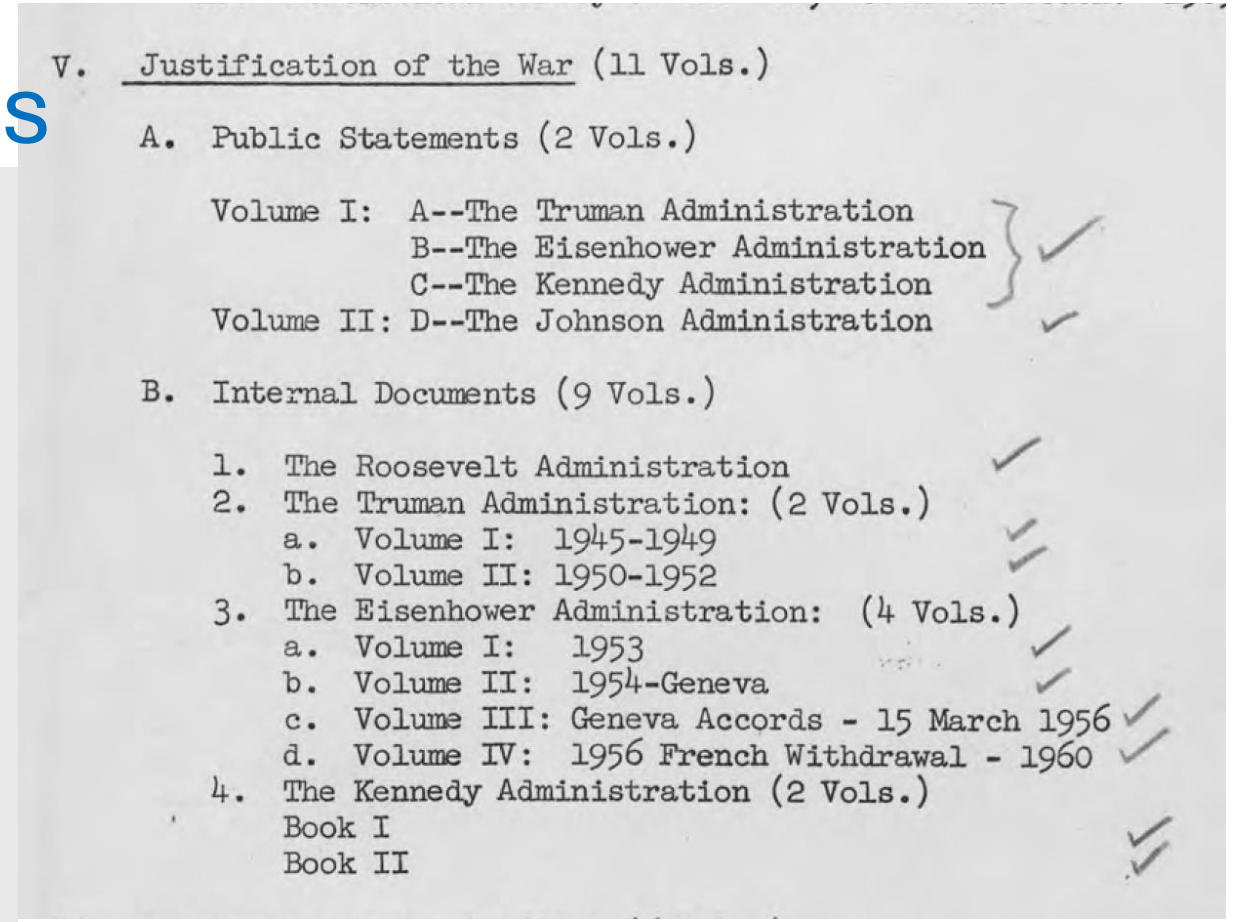
1. Trelloboards
    1. Inviteres én for hver gruppe
2. PentagonPapers
    1. På trello skal hver gruppe byde ind på et eller flere dokumenter
    2. Hver gruppe laver en kolonne med
        1. size,dates(liste),nltk-score,nounsfreq,personer,OE
        2. Upload til deres git-branch
    3. Hver gruppe undersøger deres tekst for hvad der skal graves efter

3. Github
    1. Branchworkflow
4. Opsamling
    1. Nltk - sentimentscore
    2. Spacy og language-model
        1. First tokenizer
    3. Quiz

Projektet opdateres

1. Github
    1. Pull https://github.com/cphstud/DALF25PentagonPapers
    2. Checkout branch wulf
        1. Tjek hvilke dokumenter der mangler ud fra content.txt
2. PentagonPapers – the missing files(?)
    1. Vi skal have en sammenligning af Kennedy public med internal
        1. Find de relevante original-pdf'er og OCR dem så filnavnet kan bruges til filtrering (*Pentagon-Papers-Part-V-A-Vol-IB.pdf-48.png.txt*)
        2. Upload til jeres git-branch

3. NLP
    1. Jeres text skal kunne hentes ind i Spyder så man kan se forbindelsen til indholdsfortegnelsen.
    2. Få FE:
        1. PP-document
        2. Side nummer
        3. Size
        4. Nltk – sentimentscore
        5. Spacy – most common noun pr. page



V. Justification of the War (11 Vols.)

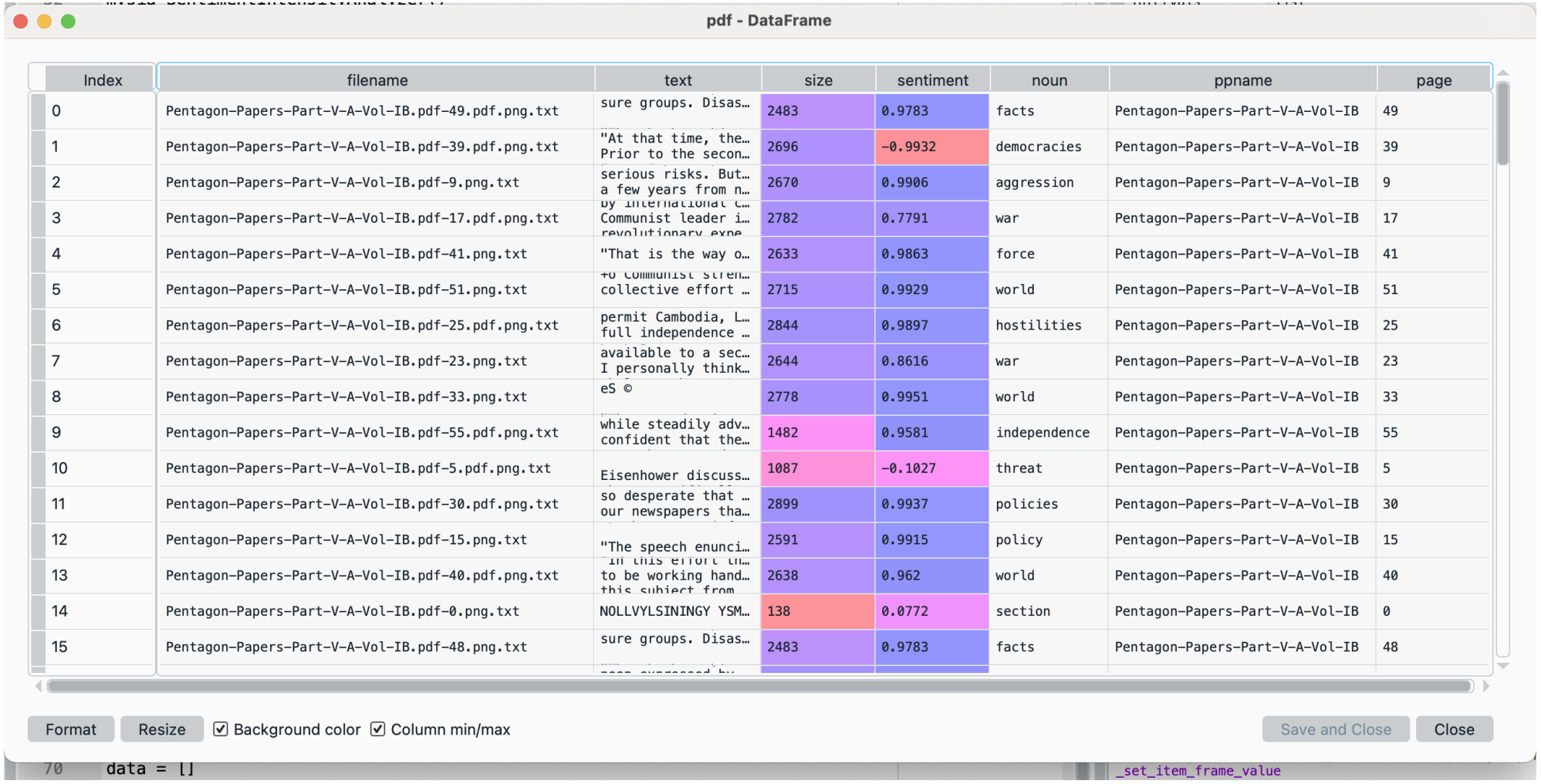
A. Public Statements (2 Vols.)

Volume I: A--The Truman Administration
B--The Eisenhower Administration
C--The Kennedy Administration
Volume II: D--The Johnson Administration

B. Internal Documents (9 Vols.)

1. The Roosevelt Administration
2. The Truman Administration: (2 Vols.)
   a. Volume I: 1945-1949
   b. Volume II: 1950-1952
3. The Eisenhower Administration: (4 Vols.)
   a. Volume I: 1953
   b. Volume II: 1954-Geneva
   c. Volume III: Geneva Accords - 15 March 1956
   d. Volume IV: 1956 French Withdrawal - 1960
4. The Kennedy Administration (2 Vols.)
   Book I
   Book II

Projektet opdateres



Index	filename	text	size	sentiment	noun	ppname	page
0	Pentagon-Papers-Part-V-A-Vol-IB.pdf-49.pdf.png.txt	sure groups. Disas…	2483	0.9783	facts	Pentagon-Papers-Part-V-A-Vol-IB	49
1	Pentagon-Papers-Part-V-A-Vol-IB.pdf-39.pdf.png.txt	"At that time, the… Prior to the secon…	2696	-0.9932	democracies	Pentagon-Papers-Part-V-A-Vol-IB	39
2	Pentagon-Papers-Part-V-A-Vol-IB.pdf-9.png.txt	serious risks. But… a few years from n…	2670	0.9906	aggression	Pentagon-Papers-Part-V-A-Vol-IB	9
3	Pentagon-Papers-Part-V-A-Vol-IB.pdf-17.pdf.png.txt	by international c… Communist leader i… revolutionary expe…	2782	0.7791	war	Pentagon-Papers-Part-V-A-Vol-IB	17
4	Pentagon-Papers-Part-V-A-Vol-IB.pdf-41.png.txt	"That is the way o…	2633	0.9863	force	Pentagon-Papers-Part-V-A-Vol-IB	41
5	Pentagon-Papers-Part-V-A-Vol-IB.pdf-51.png.txt	to communist stren… collective effort …	2715	0.9929	world	Pentagon-Papers-Part-V-A-Vol-IB	51
6	Pentagon-Papers-Part-V-A-Vol-IB.pdf-25.pdf.png.txt	permit Cambodia, L… full independence …	2844	0.9897	hostilities	Pentagon-Papers-Part-V-A-Vol-IB	25
7	Pentagon-Papers-Part-V-A-Vol-IB.pdf-23.png.txt	available to a sec… I personally think…	2644	0.8616	war	Pentagon-Papers-Part-V-A-Vol-IB	23
8	Pentagon-Papers-Part-V-A-Vol-IB.pdf-33.png.txt	eS ©	2778	0.9951	world	Pentagon-Papers-Part-V-A-Vol-IB	33
9	Pentagon-Papers-Part-V-A-Vol-IB.pdf-55.pdf.png.txt	while steadily adv… confident that the…	1482	0.9581	independence	Pentagon-Papers-Part-V-A-Vol-IB	55
10	Pentagon-Papers-Part-V-A-Vol-IB.pdf-5.pdf.png.txt	Eisenhower discuss…	1087	-0.1027	threat	Pentagon-Papers-Part-V-A-Vol-IB	5
11	Pentagon-Papers-Part-V-A-Vol-IB.pdf-30.pdf.png.txt	so desperate that … our newspapers tha…	2899	0.9937	policies	Pentagon-Papers-Part-V-A-Vol-IB	30
12	Pentagon-Papers-Part-V-A-Vol-IB.pdf-15.png.txt	"The speech enunci…	2591	0.9915	policy	Pentagon-Papers-Part-V-A-Vol-IB	15
13	Pentagon-Papers-Part-V-A-Vol-IB.pdf-40.pdf.png.txt	in this effort th… to be working hand… this subject from…	2638	0.962	world	Pentagon-Papers-Part-V-A-Vol-IB	40
14	Pentagon-Papers-Part-V-A-Vol-IB.pdf-0.png.txt	NOLLVYLSININGY YSM…	138	0.0772	section	Pentagon-Papers-Part-V-A-Vol-IB	0
15	Pentagon-Papers-Part-V-A-Vol-IB.pdf-48.png.txt	sure groups. Disas…	2483	0.9783	facts	Pentagon-Papers-Part-V-A-Vol-IB	48