

Workshop materials are at <https://bit.ly/natural-pdf>

open all these

Natural PDF basics with text and tables

Natural PDF is a spatially-aware PDF processing library



Live coding worksheet

✓ Completed version



Download: [worksheet](#) | [completed](#)



Slides: [slides.pdf](#)

Links:

click all these to install



```
!pip install "natu
```

Magical, modern PDF processing with Natural PDF

Jonathan Soma

Knight Chair in Data Journalism

Columbia University

js4571@columbia.edu / [@dangerscarf](#)

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Code	Issue	Severity	Notes
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>

Police disciplinary files (Island County Prosecuting Attorney)

Aspen Ford filed this request with the Island County Prosecuting Attorney of Island County, WA.

It is a clone of this request.

Tracking #60799

Multi RequestPolice disciplinary files

SubmittedJuly 10, 2025

STATUSCompleted

MuckRock users can file, duplicate, track, and share public records requests like this one. Learn more.

2 Communications3 Files

2025.07.11 ICPA BRADY LIST3 pages

FileAa TextPagesNotes

WASHINGTON STATE PATROL
Ahlstedt, Connor (WSP)
Alm, Mike (WSP)
Amorati, Anthony J. (WSP)
PA 000001

contributed by MuckRock Staff (MuckRock Staff)

Arnold, James
Prosecuting Attorney)

Anonymous sent this request to the Radford University of Radford, VA

Aspen Ford sent this request to the Prosecuting Attorney of Island County

Academic Dishonesty Study (Radford University)

Tom Nash filed this request on behalf of Anonymous with the Radford University of Radford, VA.

Multi RequestAcademic Dishonesty Study

SubmittedJuly 2, 2025

STATUSCompleted

MuckRock users can file, duplicate, track, and share public records requests like this one. Learn more.

3 Communications1 File

Academic Integrity Violations Fa2021-Sp2025_7.11.20256 pages

FileAa TextPagesNotes


Radford University - Academic Integrity Reports Fall 2021 - Spring 2025

Date/Time of Incident	Conduct Charges	Course
8/3/2021	Plagiarism	English 472/Shakespeare Survey
8/30/2021	Cheating, Plagiarism	Exercise, Sport and Health Education 396-01
9/30/2021	Plagiarism	English 111-78 Principles of College Composition
9/30/2021	Plagiarism	English 111-78 Principles of College Composition
10/13/2021	Cheating, Plagiarism	English 111-78 Principles of College Composition
10/13/2021	Cheating, Plagiarism	English 111-78 Principles of College Composition
10/13/2021	Cheating, Plagiarism	Anthropological Sciences
10/13/2021	Cheating	Anthropological Sciences
10/15/2021	Plagiarism	Health Communication and Coaching 2021
10/15/2021	Plagiarism	Health Communication and Coaching 2021
10/15/2021	Cheating	Communication
10/15/2021	Plagiarism	Principles of Marketing/Marketing 340
10/16/2021	Cheating	Communication
10/18/2021	Plagiarism	English 200: Literary Texts and Contexts
10/18/2021	Cheating	Polity 23101 - Organizational Biology
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3
10/19/2021	Cheating	Introductory Psychology/Psychology 121 Section 3

See all →

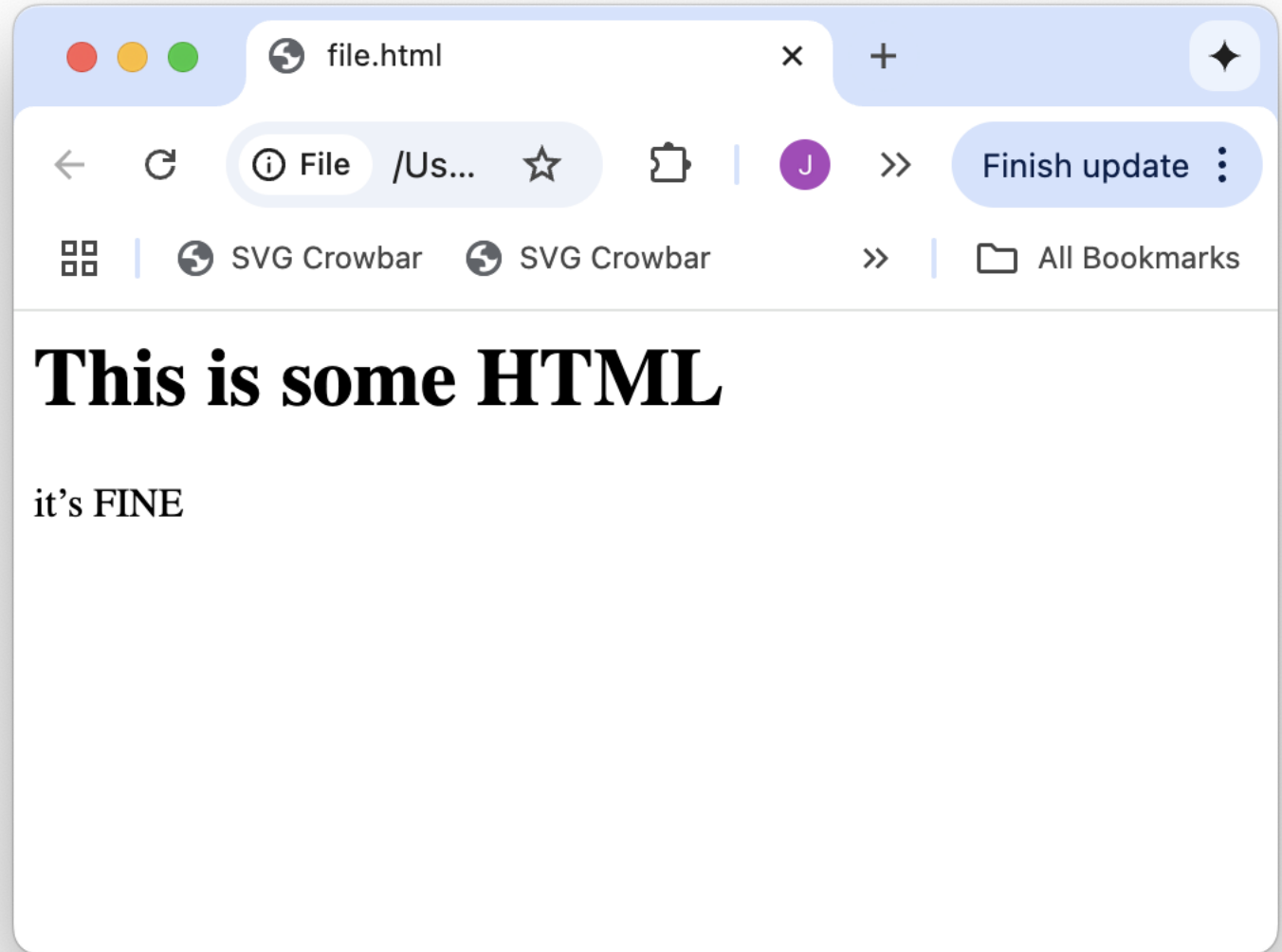
Newsletter

Want the latest investigative and FOIA news?



This is to inform you that, at the direction of President Donald J. Trump, the United States Department of Agriculture is conducting a review of its research and other

<h1>This
is some
HTML</h1>
<p>it's
FINE</h2>





This is a
PDF

it's
BAD

This is a it's

PDF

BAD

This is a

it's

PDF



BAD

matrix: (24.96, 0.0, 0.0, 24.96, 160.080072, 359.47999999999996)

fontname: AAAAAC+Georgia

adv: 0.293

upright: True

x0: 160.080072

y0: 354.01375999999993

x1: 167.393352

y1: 378.97375999999997

width: 7.313279999999992

height: 24.9600000000000036

size: 24.9600000000000036

mcid: None

tag: None

object_type: char

ncs: ICCBased

text: i

stroking_color: None

stroking_pattern: None

non_stroking_color: (0, 0, 0)

non_stroking_pattern: None



i t ' s

B A D



object_type: rect
width: 500
height: 1



This is a
PDF BAD

Radford University - Academic Integrity Reports Fall 2021 - Spring 2025

Date/Time of Incident	Conduct Charges
8/3/2021	Plagiarism
9/30/2021	Cheating, Plagiarism
9/30/2021	Plagiarism
9/30/2021	Plagiarism
9/30/2021	Plagiarism
10/11/2021	Cheating, Plagiarism
10/11/2021	Cheating, Plagiarism
10/11/2021	Cheating, Plagiarism
10/13/2021	Cheating
10/13/2021	Cheating
10/15/2021	Facilitation
10/15/2021	Plagiarism
10/15/2021	Cheating
10/15/2021	Facilitation
10/16/2021	Cheating
10/18/2021	Plagiarism
10/19/2021	Cheating
10/19/2021	Cheating
10/19/2021	Cheating
10/19/2021	Cheating
10/26/2021	Plagiarism
10/26/2021	Cheating
11/5/2021	Plagiarism
11/8/2021	Cheating, Plagiarism
11/8/2021	Facilitation
11/17/2021	Cheating
11/17/2021	Cheating
11/19/2021	Cheating, Facilitation
11/19/2021	Cheating, Plagiarism
12/3/2021	Cheating
12/3/2021	Cheating

Site: Durham's Meatpacking Chicago, Ill.
Date: February 3, 1905

Violation Count: 7

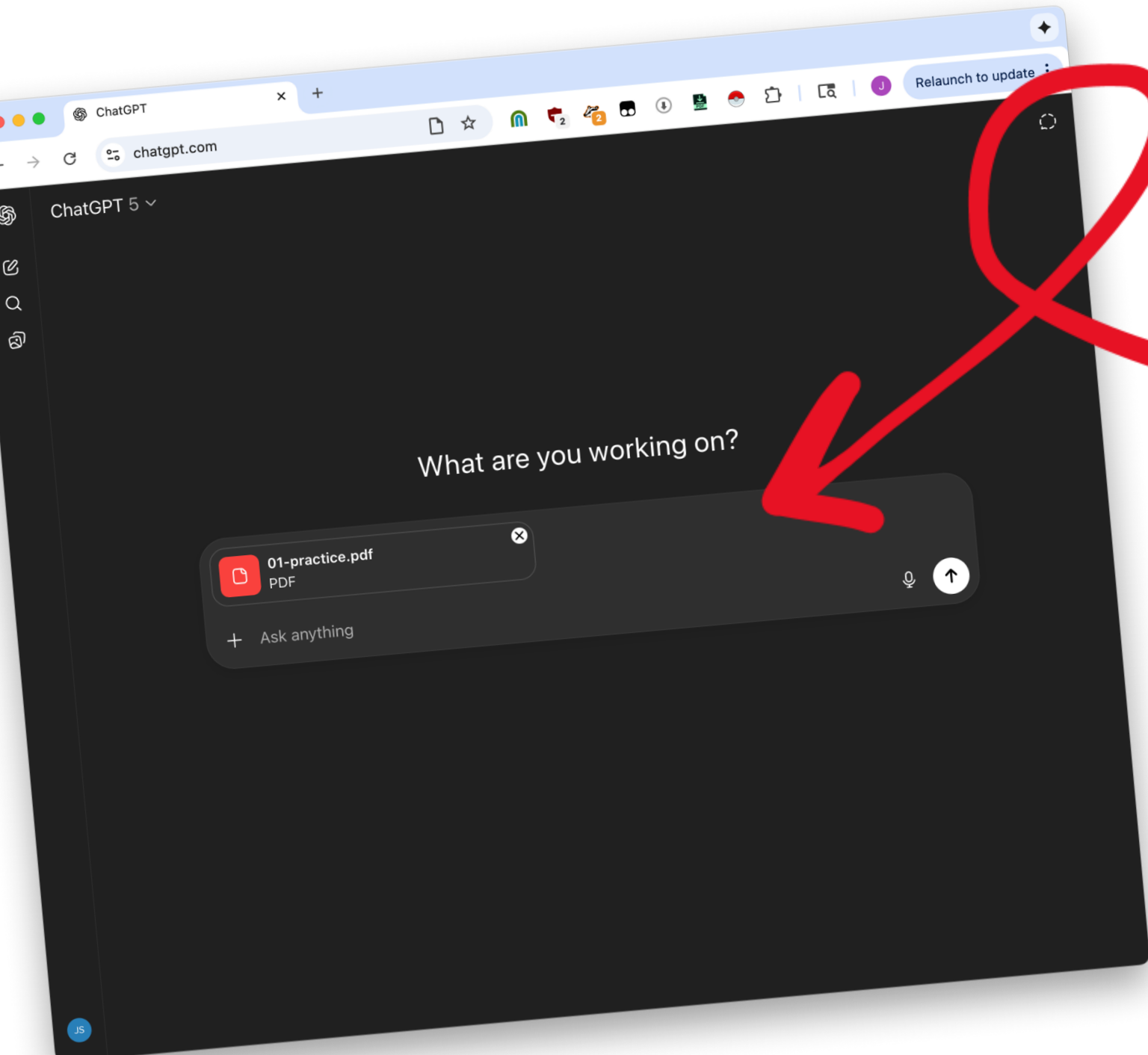
Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>





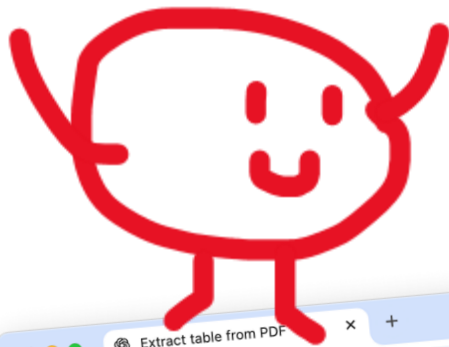
Site: Durham's Meatpacking Chicago, Ill.
Date: February 3, 1905
Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Critical	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Serious	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input checked="" type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>

Jungle Health and Safety Inspection Service



Extract table from PDF

chatgpt.com/c/68bad0d2-473c-8331-8...

Relaunch to update

Share

ChatGPT 5

01-practice.pdf PDF

extract the

Here's the table extracted from your PDF:

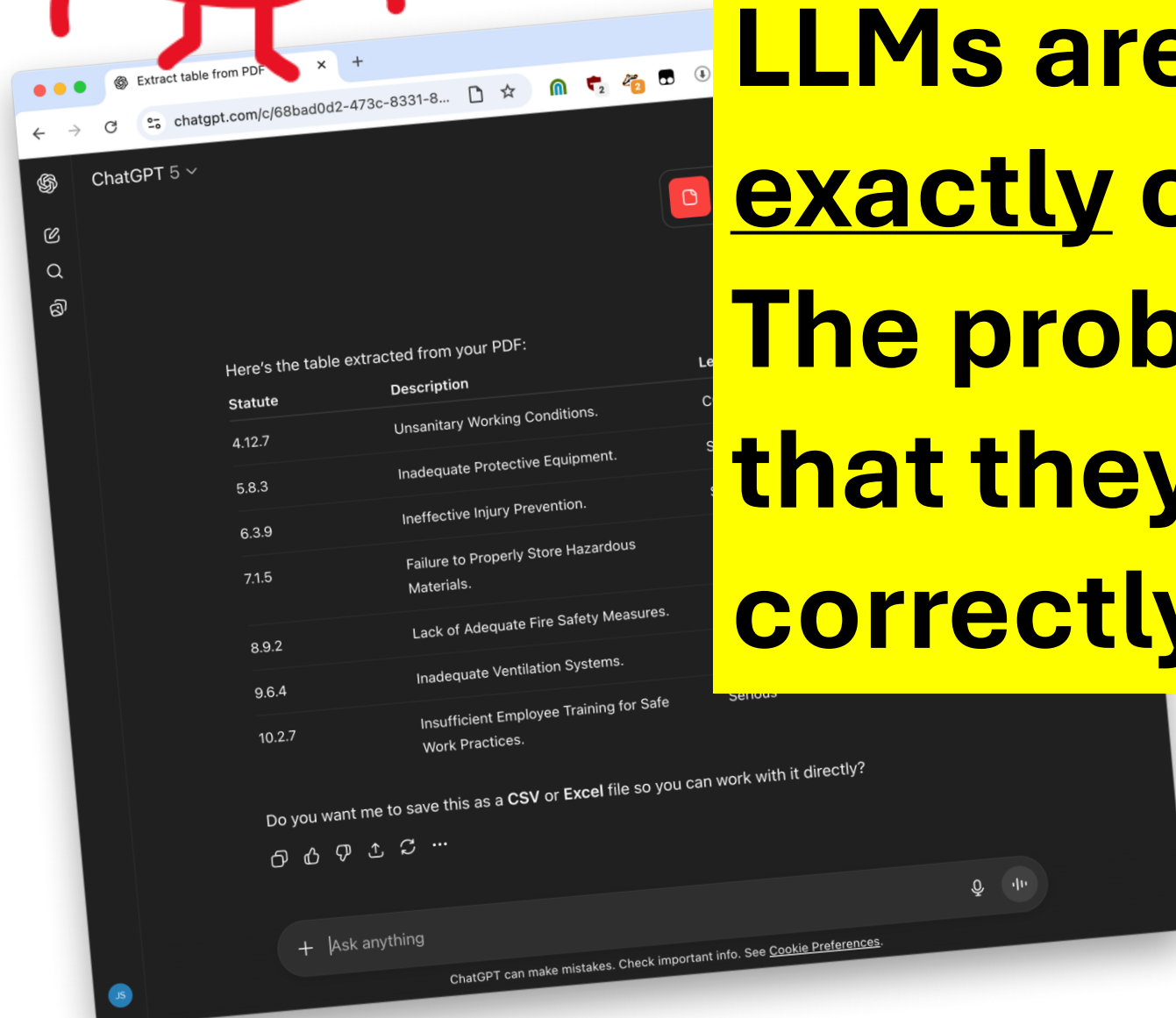
Statute	Description	Level
4.12.7	Unsanitary Working Conditions.	Critical
5.8.3	Inadequate Protective Equipment.	Serious
6.3.9	Ineffective Injury Prevention.	Serious
7.1.5	Failure to Properly Store Hazardous Materials.	Critical
8.9.2	Lack of Adequate Fire Safety Measures.	Serious
9.6.4	Inadequate Ventilation Systems.	Serious
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious

Do you want me to save this as a CSV or Excel file so you can work with it directly?

+ Ask anything

ChatGPT can make mistakes. Check important info. See [Cookie Preferences](#).

***unfortunately:* doing
this should land you
in prison, forever**



Here's the table extracted from your PDF:

Statute	Description
4.12.7	Unsanitary Working Conditions.
5.8.3	Inadequate Protective Equipment.
6.3.9	Ineffective Injury Prevention.
7.1.5	Failure to Properly Store Hazardous Materials.
8.9.2	Lack of Adequate Fire Safety Measures.
9.6.4	Inadequate Ventilation Systems.
10.2.7	Insufficient Employee Training for Safe Work Practices.

Do you want me to save this as a **CSV** or **Excel** file so you can work with it directly?



+ Ask anything

ChatGPT can make mistakes. Check important info. See [Cookie Preferences](#).

LLMs are not built for exactly copying things. The probability is high that they'll copy correctly, but not 100%.

Fund (Investment Period Beginning Date / Ending Date)	Committed Capital	Available Capital (\$)	Value	Unrealized Investments	% Public	Realized Investments	Value	Total Investments	Value	Net IRR (%)	Realized	Total
Real Estate												
BREP I (Sep 1994 / Oct 1996)	\$ 140,714	\$ -	\$ -	n/a	n/a	\$ 345,190	2.5x	\$ 345,190	2.5x	33%	33%	
BREP II (Apr 1996 / Mar 1999)	1,121,708	-	-	n/a	n/a	1,121,708	2.5x	1,121,708	2.5x	40%	40%	
BREP III (Apr 1999 / Apr 2001)	1,522,768	-	-	n/a	n/a	1,522,768	2.3x	1,522,768	2.3x	21%	21%	
BREP IV (Apr 2001 / Mar 2005)	2,531,659	-	-	n/a	n/a	2,531,659	2.4x	2,531,659	2.4x	18%	18%	
BREP V (Feb 2005 / Feb 2007)	4,684,232	-	-	n/a	n/a	4,684,232	2.5x	4,684,232	2.5x	18%	18%	
BREP VI (Aug 2007 / Apr 2011)	11,060,122	-	6,220									
BREP VII (Aug 2011 / Apr 2016)	13,505,457	1,017,261	11,827,6									
BREP VIII (Aug 2016 / Apr 2021)	16,607,961	2,006,806	11,807,9									
BREP IX (Aug 2019 / Aug 2022)	21,349,913	3,214,887	6,526,6									
BREP X (Aug 2022 / Feb 2026)	30,440,856	22,293,304	13,955,5									
Total Global BREP	106,145,290	28,532,258	47,924,2									
BREP Europe I (Sep 2001 / Sep 2005)	€ 824,172	€ -	-									
BREP Europe II (Jun 2005 / Jun 2008)	1,629,748	-	152,188									
BREP Europe III (June 2008 / Sep 2013)	3,205,420	391,552	4,467,44									
BREP Europe IV (July 2013 / Oct 2016)	4,676,577	1,079,242	4,467,44									
BREP Europe V (Oct 2016 / Oct 2019)	7,962,703	884,823	4,607,44									
BREP Europe VI (Oct 2019 / Sep 2023)	9,933,950	2,195,177	4,467,440	1.1x	n/a	5,440,555	1.3x	9,908,005	2.2x	n/a	6%	
*BREP Europe VII (Sep 2023 / Mar 2029)	7,750,000	2,204,643	1,768,948	1.0x	26	-	-	1,768,948	1.0x	n/a	n/a	
Total BREP Europe	35,982,570	6,755,437	15,853,456	1.2x	26	36,477,174	2.4x	52,330,630	1.8x	16%	13%	
BREP Asia I (Dec 2013 / Dec 2017)	4,262,073	896,533	1,540,757	1.2x	14%	7,229,527	2.4x	8,770,284	2.8x	16%	12%	
BREP Asia II (Dec 2017 / Mar 2022)	7,254,811	1,310,779	1,540,757	1.1x	n/a	2,112,419	1.6x	3,653,176	2.7x	n/a	5%	
BREP Asia III (Dec 2022 / Sep 2027)	-	-	-									
Total BREP Asia	-	-	-									
*Total BREP International	-	-	-									
*BRED High-Yield (Various) (N)	-	-	-									

numbers + structured data require accuracy

you get a table but probably not the table

Ground truth

REVENUE, INCOME AND EXPENSES

in €m	Jan - Sep 2024	Jan - Sep 2023	Change in %
Traffic revenue	23,578	22,583	4
Other revenue	4,559	4,098	11
Total revenue	28,137	26,681	5
Other operating income			

of which other raw materials, consumables and supplies and purchased services	3,796	3,360	
of which fees and charges	1,942	1,594	
of which external MRO services	6,021	5,667	
Staff costs	1,870	1,805	
	3,765	3,706	
	29,140	26,571	
	138	105	
equity investments	1,177	2,280	
	72	-62	
BIT	1,249	2,218	
	-209	-248	
	-54	50	
	986	2,020	
income taxes	-130	-362	
	855	1,654	
	-15	-48	
	840	1,606	
	10	0	
	830	1,606	

Profit/loss from continuing operations	-10	-48
Profit/loss from discontinued operations		
Profit/loss after income taxes		
Profit/loss attributable to minority interests		
Net profit/loss attributable to shareholders of Deutsche Lufthansa AG	830	1,606
Profit/loss to minority		
Net profit/loss attributable to shareholders of Deutsche Lufthansa AG	830	1,606

Andrew Ng OCR Output

www.runpulse.com/blog

Workshop materials are at <https://bit.ly/natural-pdf>

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

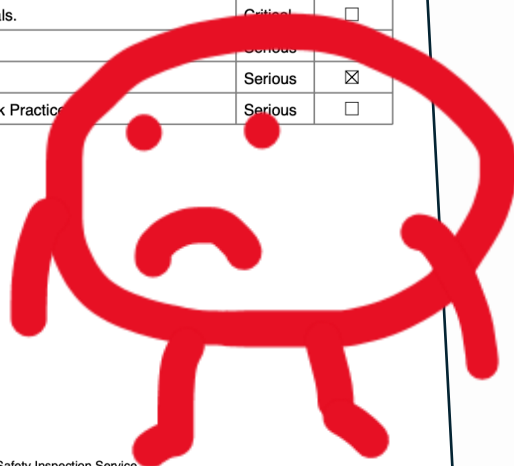
Summary: Worst of any, however, were the fertilizer men, and those who These people could not be shown to the visitor - for the odor of a fertilizer visitor at a hundred yards, and as for the other men, who worked in tank some of which there were open vats near the level of the floor, their peculiar into the vats; and when they were fished out, there was never enough of exhibiting - sometimes they would be overlooked for days, till all but the bones to the world as Durham's Pure Leaf Lard!

sometimes, you really do
need to do it manually

INS-UP70N51NCL41R

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input checked="" type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>



Jungle Health and Safety Inspection Service

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!



Workshop materials are at <https://bit.ly/natural-pdf>

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Workshop materials are at <https://bit.ly/natural-pdf>



Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Workshop materials are at <https://bit.ly/natural-pdf>

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>

Workshop materials are at <https://bit.ly/natural-pdf>

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input checked="" type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>

Workshop materials are at <https://bit.ly/natural-pdf>

open all these

Your notebooks

These are where all of the action happens! Click them to enter.

- [01 - Natural PDF basics with text and tables](#)
- [02 - OCR and AI magic](#)
- [03 - AI and data extraction](#)
- [04 - Page structure](#)
- and as a special bonus.. [05 - Advanced PDF manipulation](#)

click all these to install

```
!pip install "natu
```

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input checked="" type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>

OCR: when text isn't text

Jungle Health and
INS-UP70N51NCL

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking
These people could not be shown to the visitor - for the odor of a fertilizer man would scare
visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam
some of which there were open vats near the level of the floor, their peculiar trouble was that
into the vats; and when they were fished out, there was never enough of them left to be wor
exhibiting - sometimes they would be overlooked for days, till all but the bones of them had
to the world as Durham's Pure Leaf Lard!

OCR: when text isn't text

```
import cv2
import pytesseract

img_cv = cv2.imread('filename.png')

img_rgb = cv2.cvtColor(img_cv, cv2.COLOR_BGR2RGB)
pytesseract.image_to_string(img_rgb)
```

```
import easyocr
reader = easyocr.Reader(['en', 'ch_sim'])
result = reader.readtext('filename.jpg')
```

```
from PIL import Image
from surya.foundation import FoundationPredictor
from surya.recognition import RecognitionPredictor
from surya.detection import DetectionPredictor

image = Image.open(IMAGE_PATH)
foundation_predictor = FoundationPredictor()
recognition_predictor = RecognitionPredictor(foundation_predictor)
detection_predictor = DetectionPredictor()

predictions = recognition_predictor([image], det_predictor=detection_predictor)
```

```
from paddleocr import PaddleOCR

ocr = PaddleOCR(
    use_doc_orientation_classify=False,
    use_doc_unwarping=False,
    use_textline_orientation=False)

result = ocr.predict("./general_ocr_002.png")
for res in result:
    res.print()
```

`.apply_ocr` various flavors of image-to-text

```
pdf.apply_ocr('surya')
```

```
pdf.apply_ocr('easyocr')
```

```
pdf.apply_ocr('paddle')
```

Workshop materials are at <https://bit.ly/natural-pdf>

every battery included

Main Tasks Libraries Languages Licenses
Other

Filter

Modal

Audio-Text-to-Text

Image-Text-to-Text

Visual Question Answering

Document Question Answering

Video-Text-to-Text

Visual Document Retrieval

Any-to-Any

Computer Vision

Depth Estimation

Object Detection

Text-to-Image

Image-to-Image

Unconditional Image Generation

Video Classification

Zero-Shot Image Classification

Mask Generation

Zero-Shot Object Detection

Image-to-3D

Keypoint Detection

Image Classification

Image Segmentation

Image-to-Text

Image-to-Video

Text-to-Video

Text-to-3D

Image Feature Extraction

Video-to-Video

Search models, datasets, users...

Models 2,046,200

Models

Datasets

Spaces

Community

Relaunch

Microsoft/Voice-1.5B

Text-to-Speech • 3B • Updated 5 days ago • 203k • 1.5k

tencent/HunyuanWorld-Voyager

Image-to-Video • Updated 1 day ago • 256 • 452

meituan-longcat/LongCat-Flash-Chat

Text Generation • 562B • Updated 6 days ago • 20k • 411

google/embeddinggemma-3.0m

Sentence Similarity • 0.3B • Updated 1 day ago • 3.87k • 287

swiss-ai/Apertus-8B-Instruct-2509

Text Generation • 8B • Updated about 6 hours ago • 19.9k • 225

moonshotai/Kimi-K2-Instruct-0905

Text Generation • Updated 1 day ago • 97 • 218

apple/FastVLM-0.5B

Text Generation • 0.8B • Updated 3 days ago

Workshop materials are at <https://bit.ly/natural-pdf>

`.ask` extractive q&a

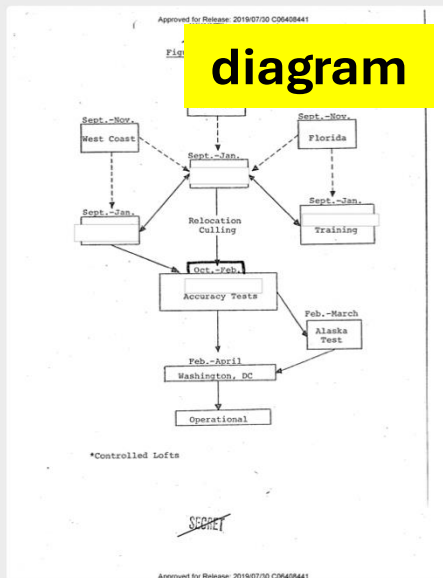
`.classify` text or vision classification

`.apply_ocr` various flavors of image-to-text

`.judge` custom region classifiers

`.correct_ocr` traditional LLMs

`.analyze_layout` document layout analysis



5



Approved for Release: 2019/07/30 C06408441

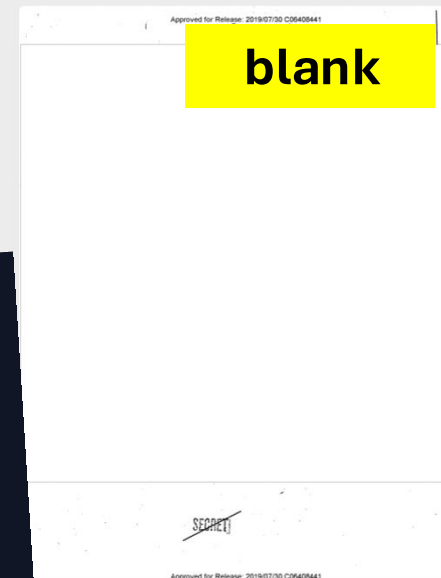
text

receiver to assist in
carried cameras will al
in some cases. Figure
A command post
communication for coordinating the field tests.
Table 1 shows the expected transitional quarter costs
(\$108,000), and Table 2 shows the expected costs for FY-77
as \$135,000 of which \$70,000 is for QWD and \$65,000 is for
OTS. Neither office has currently budgeted for these FY-77
costs.

OTS SUPPORT
During the T.O. two basic activities will take place:
A pigeon loft and facilities will be constructed

~~SECRET~~

Approved for Release: 2019/07/30 C06408441



8

```
pdf.classify_pages(  
    ['diagram', 'text', 'invoice', 'blank'],  
    using='vision'  
)
```

Approved for Release: 2019/07/30 C06408441

invoice

CONTRACTOR	ITEM COST \$	SUB TOTAL	RUNNING TOTAL
Materials & Supplies (Birds, Lofts/Training)	500		
Purchase of Birds	3,000		
Aircraft Rental \$15/hr (100 hrs.)	1,500		
Labor	3,000		
Travel	2,000	10,000	10,000
5 Cameras at \$2,000 each	10,000		
Purchase of Birds	3,000		
Labor and Repair	7,000		
Travel and Field Support	5,000	25,000	35,000
4 DF Receivers at \$825 each	3,300		
Eight 1-gr. Transmitters at \$50 each	400		
Three 6-gr. Transmitters at \$65 each	200		
Commo Equipment for Field Tests	3,000		
Refit of Aircraft	1,600		
Labor	4,500		
Travel	3,000	16,000	51,000

(continued)

~~SECRET~~

Approved for Release: 2019/07/30 C06408441

9

Approved for Release: 2019/07/30 C06408441

invoice

CONTRACTOR	ITEM COST \$	SUB TOTAL	RUNNING TOTAL
Drawings & Doc. (New Camera)	4,000		
Purchase Birds & West Coast	4,000		
Labor	7,000		
Travel	5,000	20,000	71,000
QRC Support	4,000	4,000	4,000
Building Supplies	2,300		
Labor	700	3,000	78,000
Specialized Relocation Training		30,000	108,000

already omitted need not be restated.

not

long ago over

~~SECRET~~

Approved for Release: 2019/07/30 C06408441

10

Approved for Release: 2019/07/30 C06408441

invoice

CONTRACTOR	ITEM COST \$	SUB TOTAL	RUNNING TOTAL
Building Supplies	500		
Feed	500		
Two Bird Handlers	21,000		
Transport, Shipping, Misc.	1,000		
Labor	1,000		
Boat Rental	1,000	25,000	25,000
QRC Support	38,000	38,000	63,000
ELMENDORF AFB (ALASKA)			
Building Supplies	1,000		
Labor	4,000		
Gen. Support	2,000	7,000	70,000

~~SECRET~~

Approved for Release: 2019/07/30 C06408441

11

Approved for Release: 2019/07/30 C06408441

invoice

CONTRACTOR	ITEM COST \$	SUB TOTAL	RUNNING TOTAL
10 + 20' Two stage effort			
30 Cameras (\$2,000 each)	60,000	60,000	60,000
Construction of Lofts and D.C. Test Support	5,000	5,000	65,000
TOTAL FY-77 COSTS			\$ 135,000

~~SECRET~~

Approved for Release: 2019/07/30 C06408441

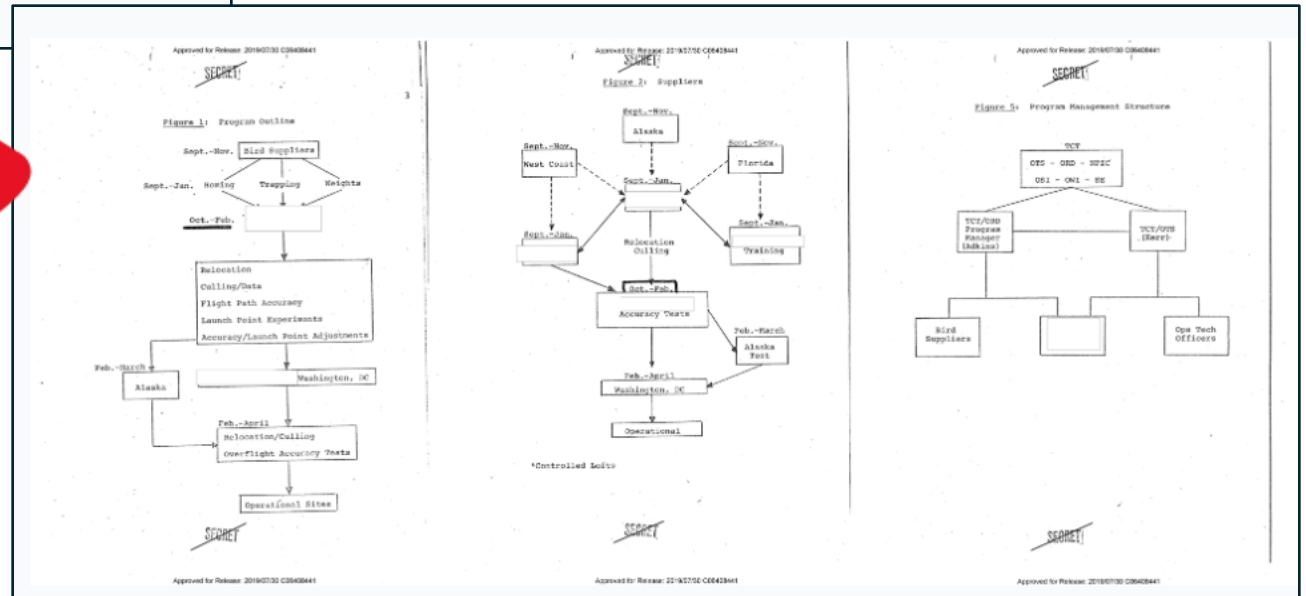
12

```
groups = pdf.pages.groupby(lambda page: page.category)
groups.info()
```

PageGroupBy with 3 groups:

```
[0] 'text': 7 pages
[1] 'diagram': 3 pages
[2] 'invoice': 7 pages
```

```
diagrams = groups.get('diagram')
diagrams.show()
```



Workshop materials are at <https://bit.ly/natural-pdf>

Jungle Health and Safety Inspection Service
INS-UP70N51NCL41R

Site: Durham's Meatpacking Chicago, Ill.

Date: February 3, 1905

Violation Count: 7

Summary: Worst of any, however, were the fertilizer men, and those who served in the cooking rooms. These people could not be shown to the visitor - for the odor of a fertilizer man would scare any ordinary visitor at a hundred yards, and as for the other men, who worked in tank rooms full of steam, and in some of which there were open vats near the level of the floor, their peculiar trouble was that they fell into the vats; and when they were fished out, there was never enough of them left to be worth exhibiting - sometimes they would be overlooked for days, till all but the bones of them had gone out to the world as Durham's Pure Leaf Lard!

Violations

Statute	Description	Level	Repeat?
4.12.7	Unsanitary Working Conditions.	Critical	<input checked="" type="checkbox"/>
5.8.3	Inadequate Protective Equipment.	Serious	<input checked="" type="checkbox"/>
6.3.9	Ineffective Injury Prevention.	Serious	<input type="checkbox"/>
7.1.5	Failure to Properly Store Hazardous Materials.	Critical	<input type="checkbox"/>
8.9.2	Lack of Adequate Fire Safety Measures.	Serious	<input type="checkbox"/>
9.6.4	Inadequate Ventilation Systems.	Serious	<input checked="" type="checkbox"/>
10.2.7	Insufficient Employee Training for Safe Work Practices.	Serious	<input type="checkbox"/>

Magical, modern PDF processing with Natural PDF

Jonathan Soma

Knight Chair in Data Journalism

Columbia University

js4571@columbia.edu / @dangerscarf