Redaction Project

Carrie Hay Kellar and Mia Bramel, Summer 2020

Agenda

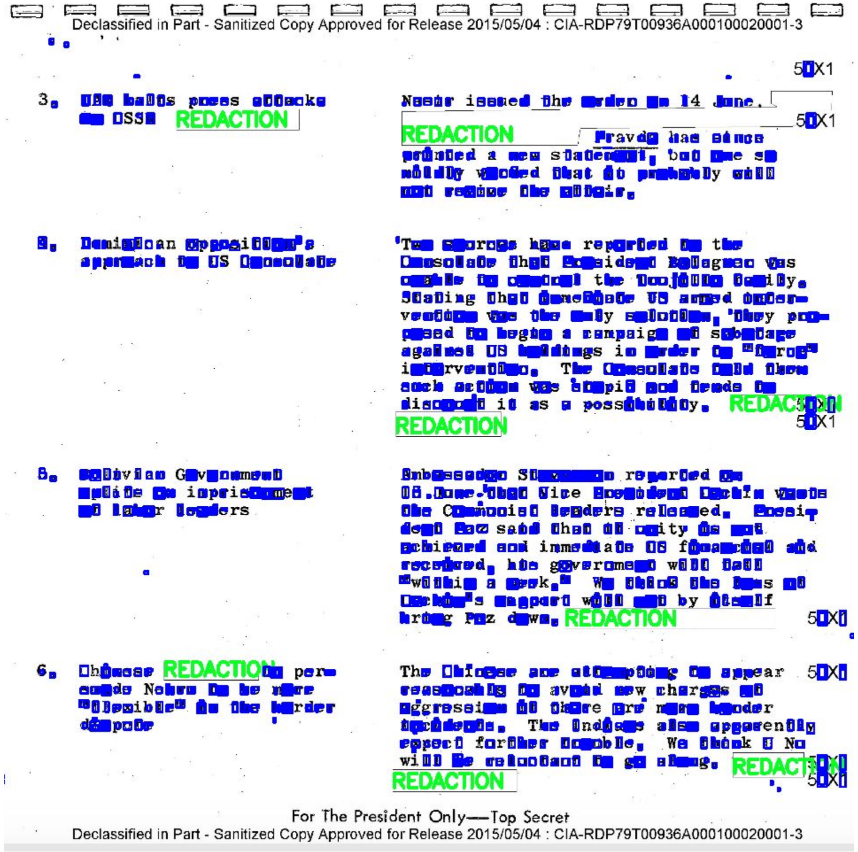
1. Method
2. Accuracy
3. Results
   a. Year
   b. NER
   c. Topic Modeling
4. Next Steps

Method

Redaction Finder

1. Determine contours of page, find bounding box
2. If bounding box is 95% white pixels, and a certain size then is potentially a redaction
3. Return largest overlapping redaction by determining intersection over union score

Redacted Text Percentage



Declassified in Part - Sanitized Copy Approved for Release 2015/05/04 : CIA-RDP79T00936A000100020001-3

For The President Only—Top Secret

Declassified in Part - Sanitized Copy Approved for Release 2015/05/04 : CIA-RDP79T00936A000100020001-3

REDACTION
REDACTION
REDACTION
REDACTION
REDACTION
REDACTION
REDACTION
REDACTION
REDACTION

- Find the percentage of the redaction rectangle which originally contained text. We do this by calculating the area of the text and divide it by the total area of the bounding rectangle.

- On average, we estimated that 26% of the bounding rectangle contains text.

- Now we can estimate the total number of words redacted on the page.

- For example, on this page, we calculated that approximately 9% of the text was redacted.

Accuracy

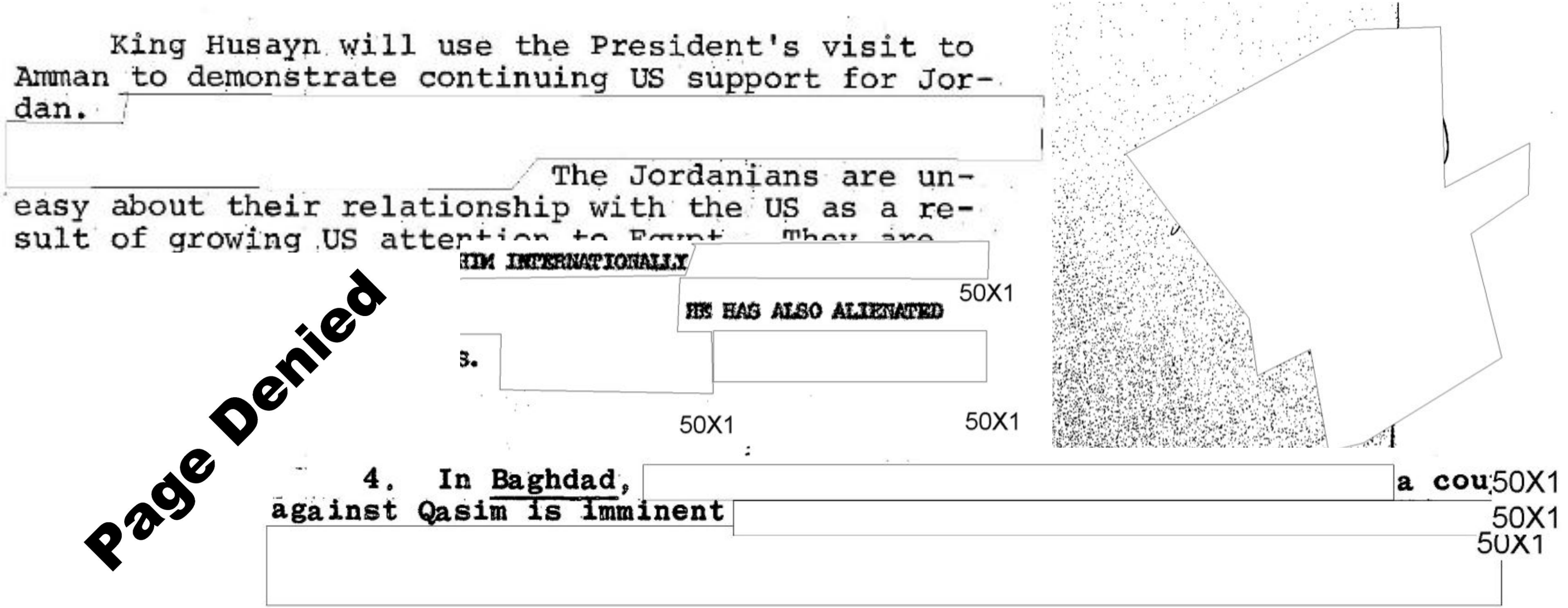
Test Case Metrics

- 200 file test case, roughly 4% of PDBs

Method	Average Redactions per Document	Accuracy
Test Case	20.33	100%
OCR with redaction code	15.57	76.5%
Our script	18.78	92.8%

- 16.3% accuracy improvement

Troubleshooting



King Husayn will use the President's visit to Amman to demonstrate continuing US support for Jordan. [redacted] The Jordanians are uneasy about their relationship with the US as a result of growing US attention to Egypt. They are

HIM INTERNATIONALLY [redacted] 50X1

HE HAS ALSO ALIENATED

S. [redacted] [redacted]

50X1 50X1

[redacted]

Page Denied

4. In Baghdad, [redacted] a cou 50X1
against Qasim is imminent [redacted] 50X1
50X1
[redacted]

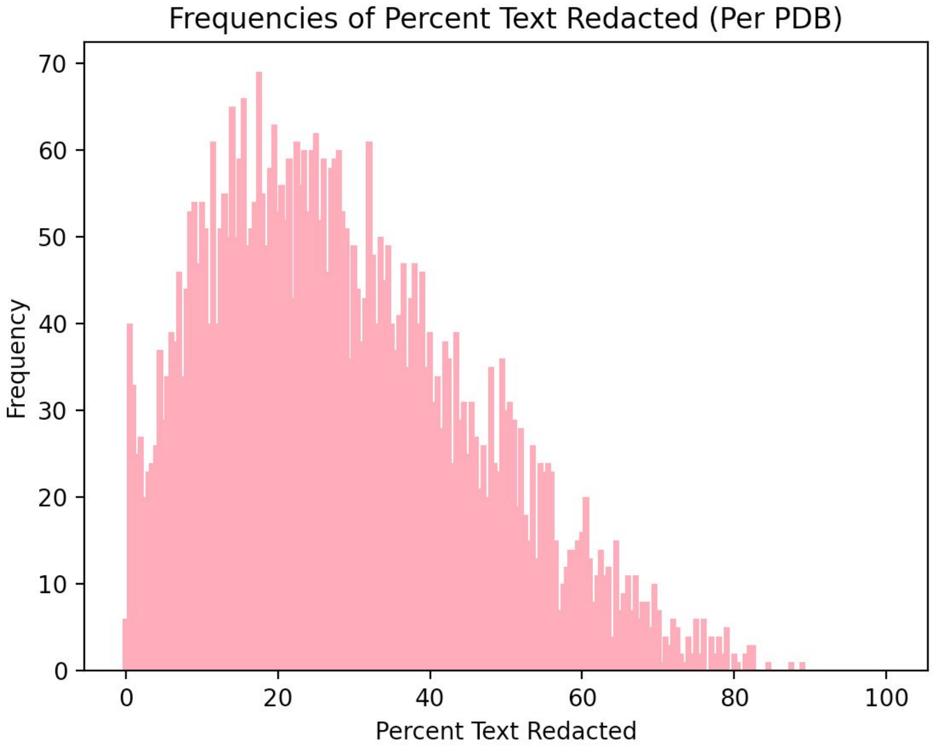
Results

Aggregate

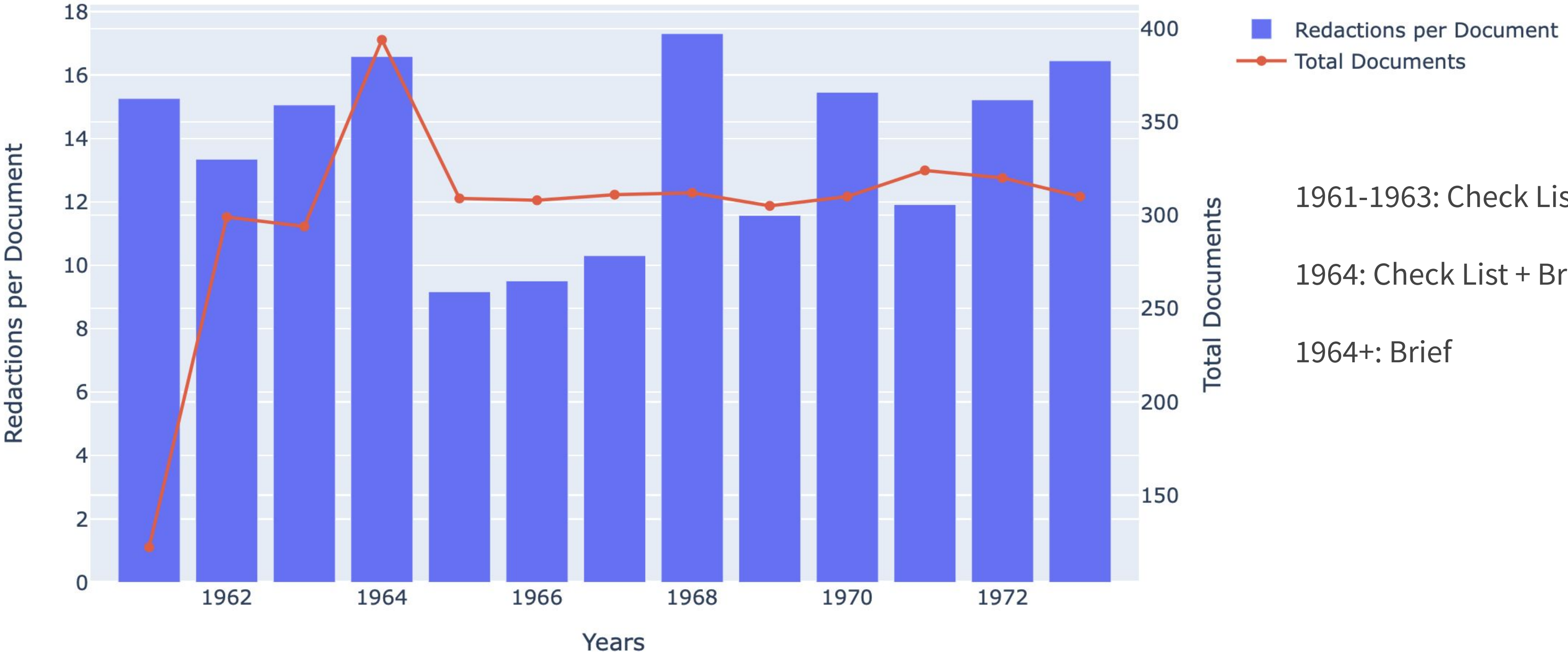
Average Redaction Count: 13

Average Percent of Text Redacted: 29%

Average Number of Words Redacted: 315



Weighted Redactions per Year



1961-1963: Check List

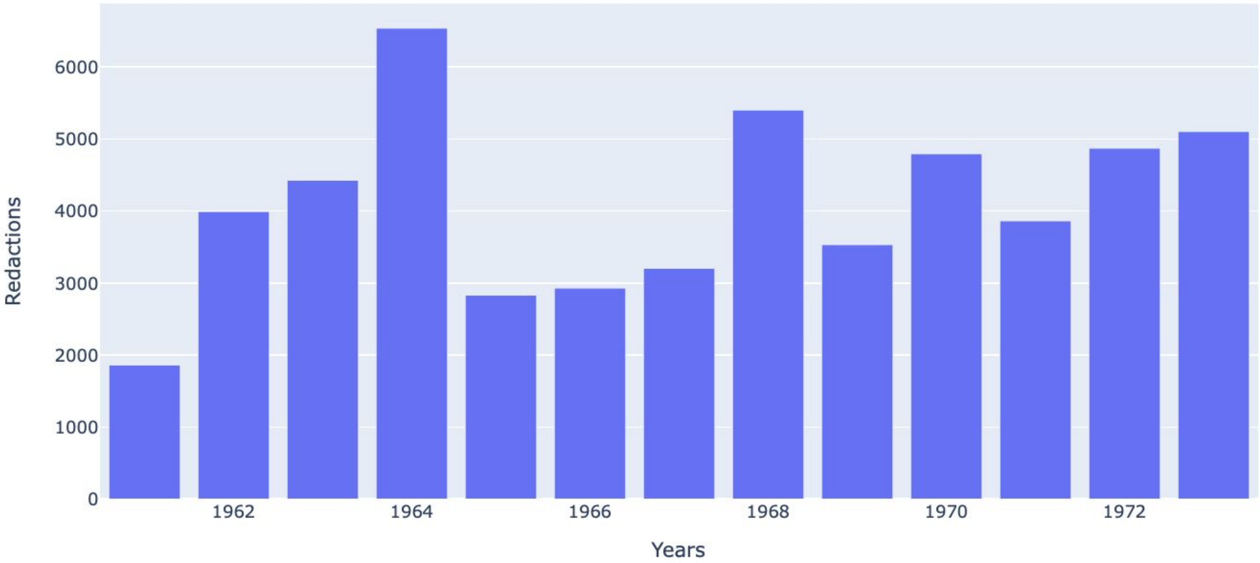
1964: Check List + Brief

1964+: Brief

Redactions per Year

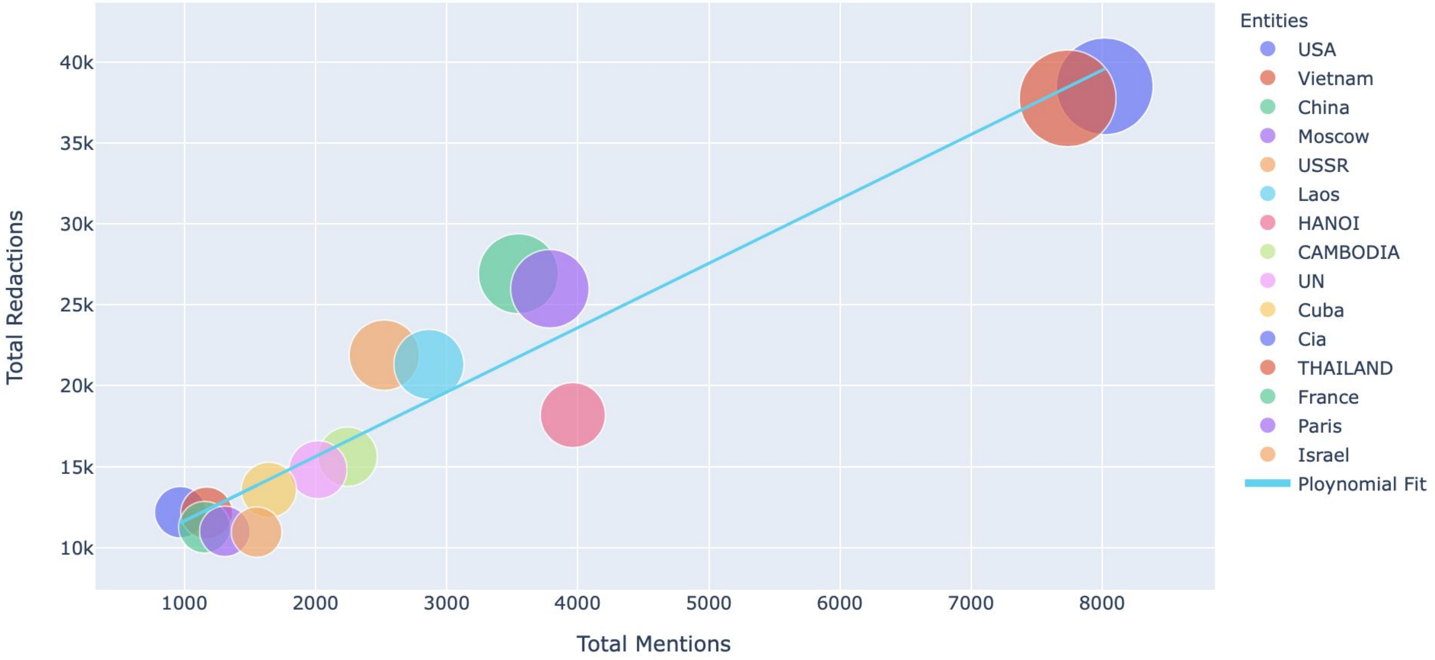
1964: Checklist seems to be more heavily redacted than Brief, smaller redactions

Majority of Cyprus documents in 1964



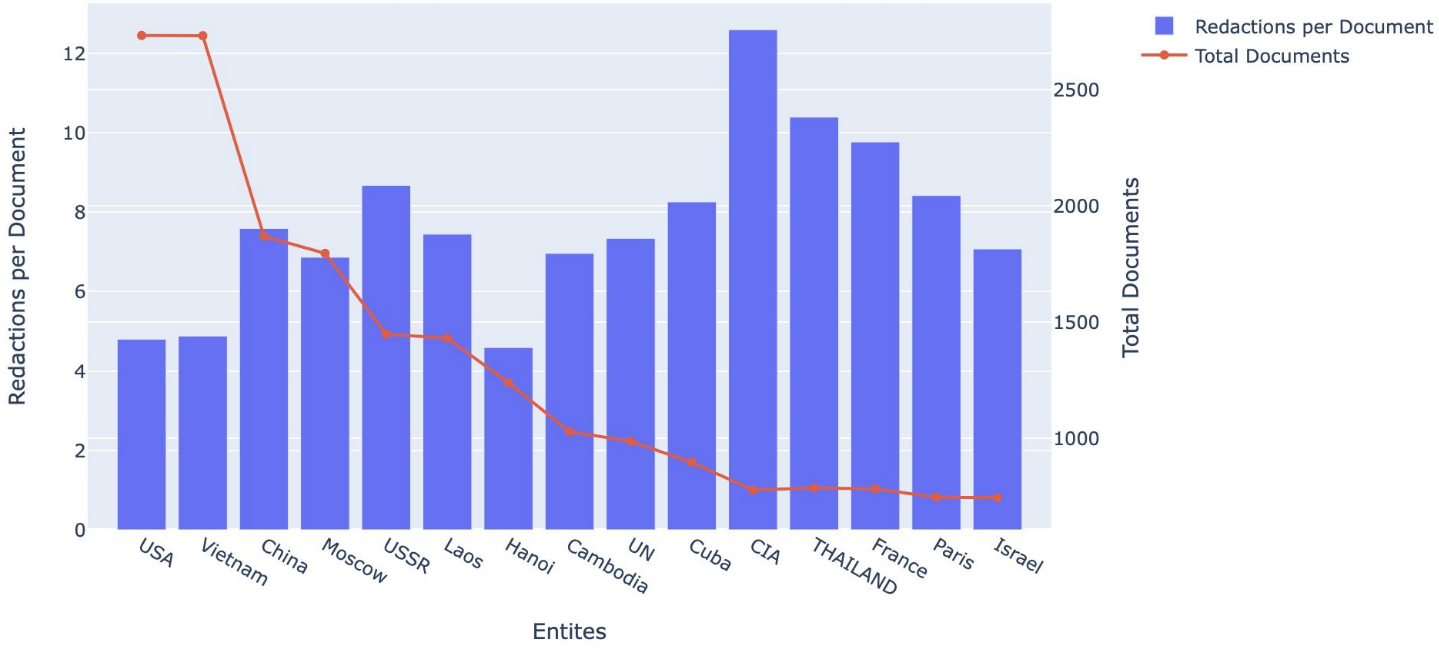
NER Top 15 Entities - Redactions per Mention

China, Moscow, USSR more heavily redacted. Hanoi less redacted.



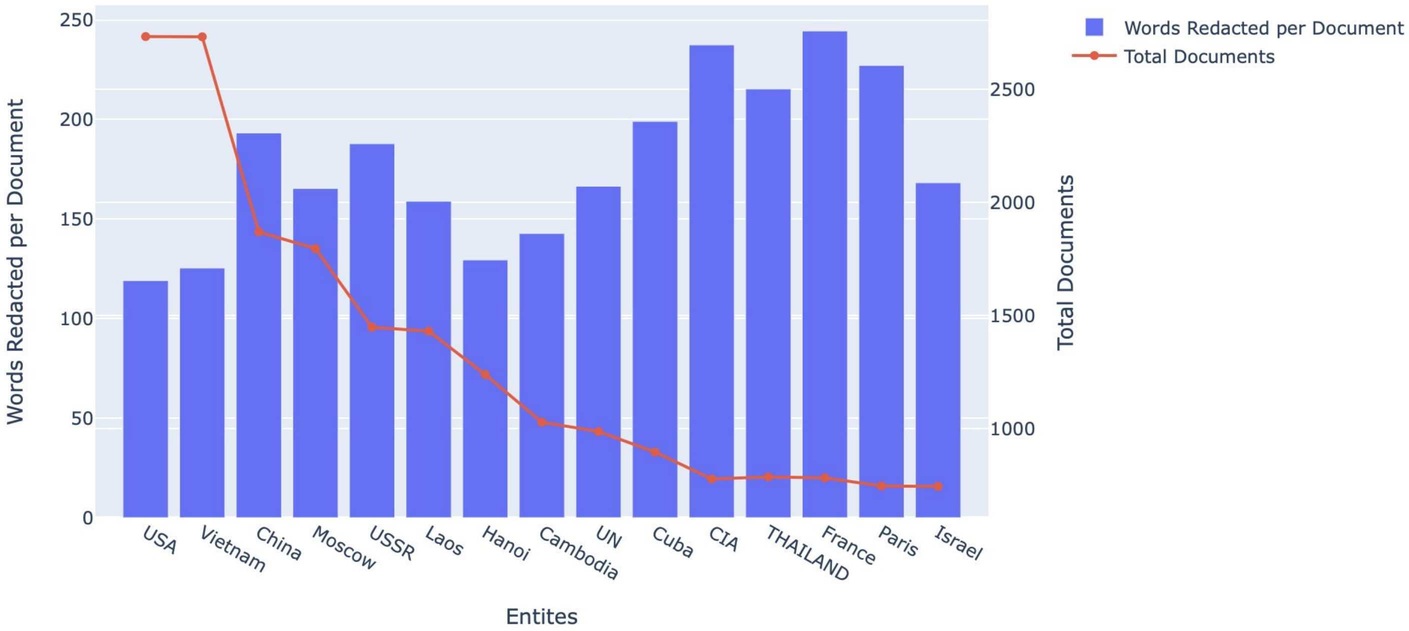
NER Top 15 Entities - Weighted Redactions per Document

Large amount of documents for USA and Vietnam: Vietnam Special Reports, USA mentions



NER Top 15 Entities - Weighted Words Redacted per Document

See a more equal distribution than redactions per document

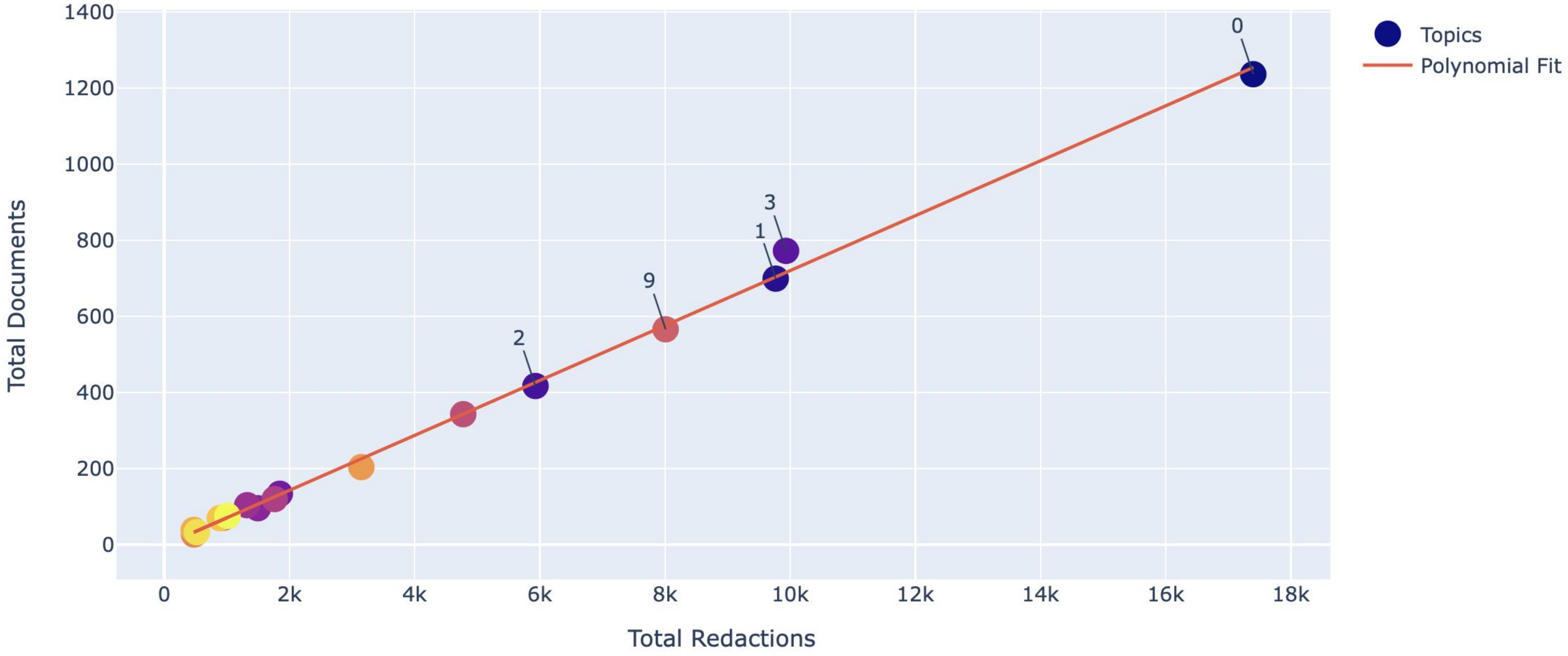


Top 5 Redacted Topics

```
(0,
  '0.009*"cont" + 0.007*"cuba" + 0.006*"sox" + 0.006*"lao" + 0.005*"chinese" + 0.005*"eee" + 0.004*"berlin" + 0.004*"tha"'),
 (1,
  '0.016*"cambodia" + 0.012*"phnom" + 0.011*"penh" + 0.011*"province" + 0.011*"phnom_penh" + 0.010*"south_vietnamese" +
0.010*"enemy" + 0.010*"cambodian"'),
 (2,
  '0.009*"portugal" + 0.009*"korea" + 0.009*"prime" + 0.008*"korean" + 0.008*"prime_minister" + 0.006*"movement" +
0.006*"portuguese" + 0.005*"cabinet"'),
 (3,
  '0.006*"student" + 0.005*"demonstration" + 0.005*"election" + 0.005*"dominican" + 0.005*"strike" + 0.004*"union" + 0.004*"embassy" +
0.004*"group"')
(9,
  '0.007*"german" + 0.006*"election" + 0.006*"germany" + 0.005*"policy" + 0.005*"relation" + 0.004*"peking" + 0.004*"brezhnev" +
0.004*"chinese"')
(11,
  '0.033*"million" + 0.032*"ton" + 0.025*"grain" + 0.020*"lon" + 0.016*"crop" + 0.015*"harvest" + 0.015*"million_ton" + 0.015*"nol"')
```

Topics - Redactions per Document

Topic modeling difficult for PDBs due to number of topics per document. Looked at redactions not by page, but by document

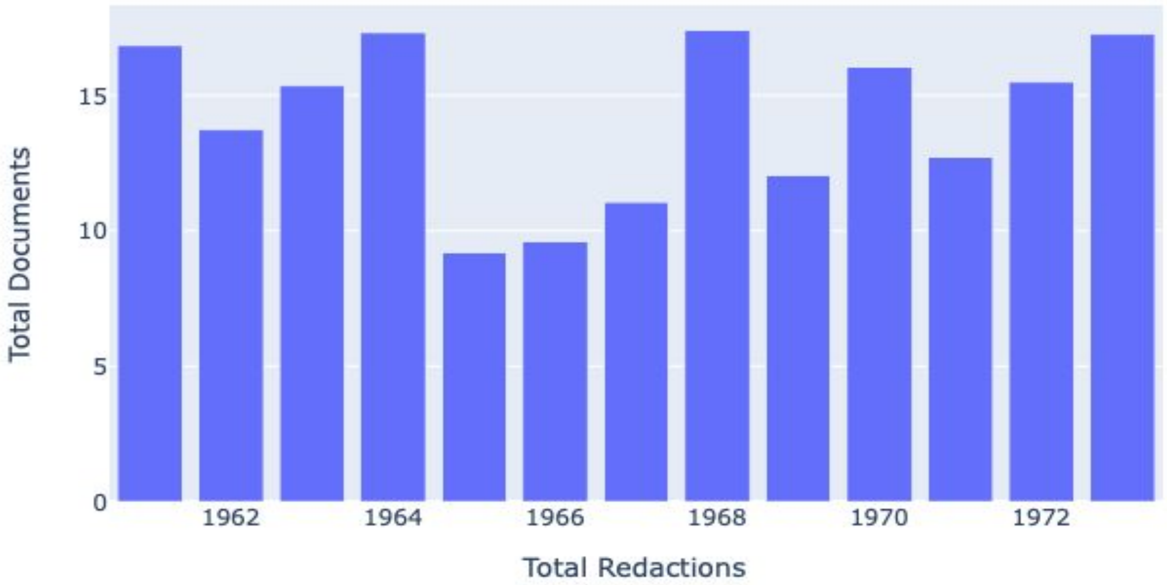


Next Steps

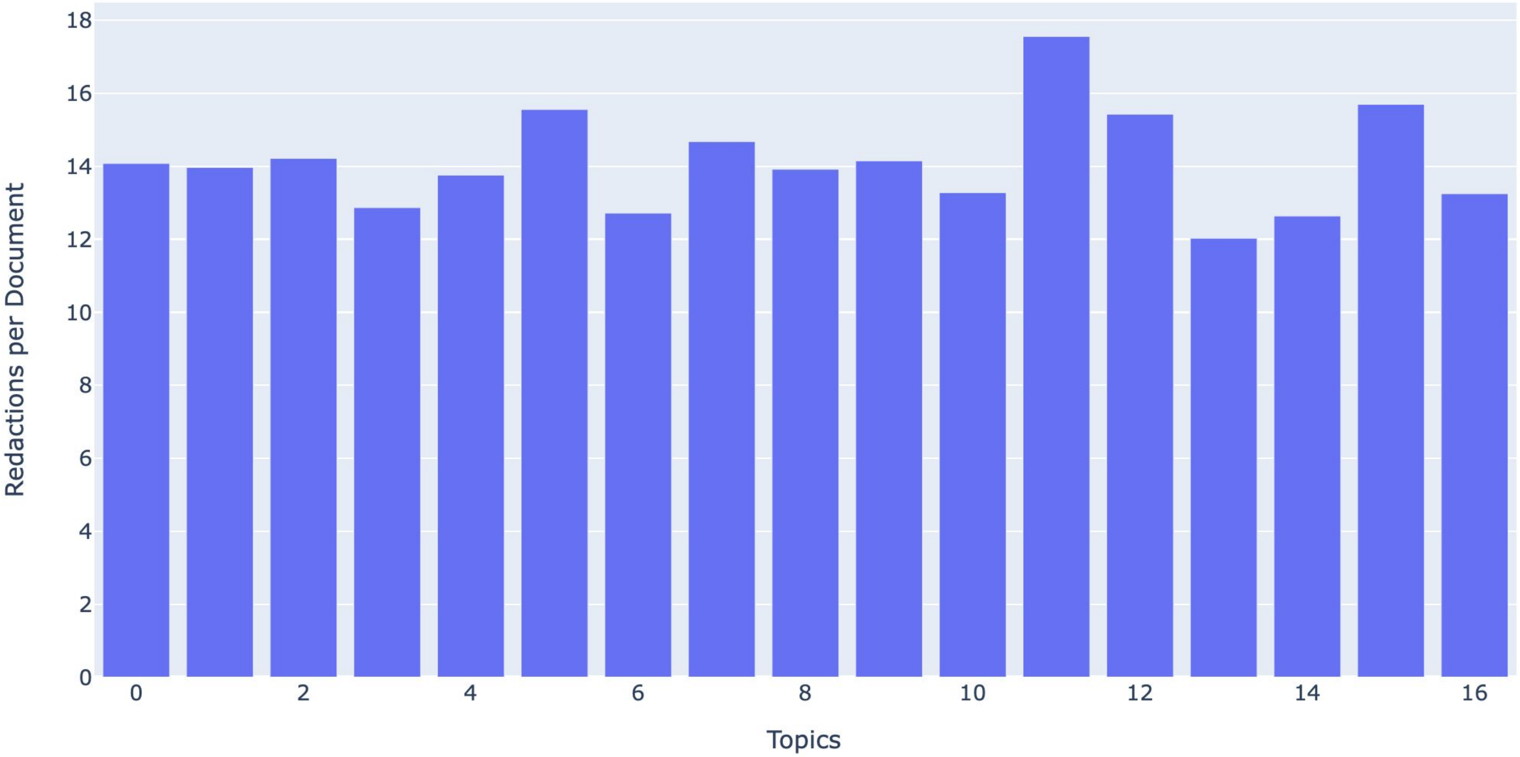
- ML / Object Recognition → for maps, redaction boxes
- Generalize for other document types?

Appendix

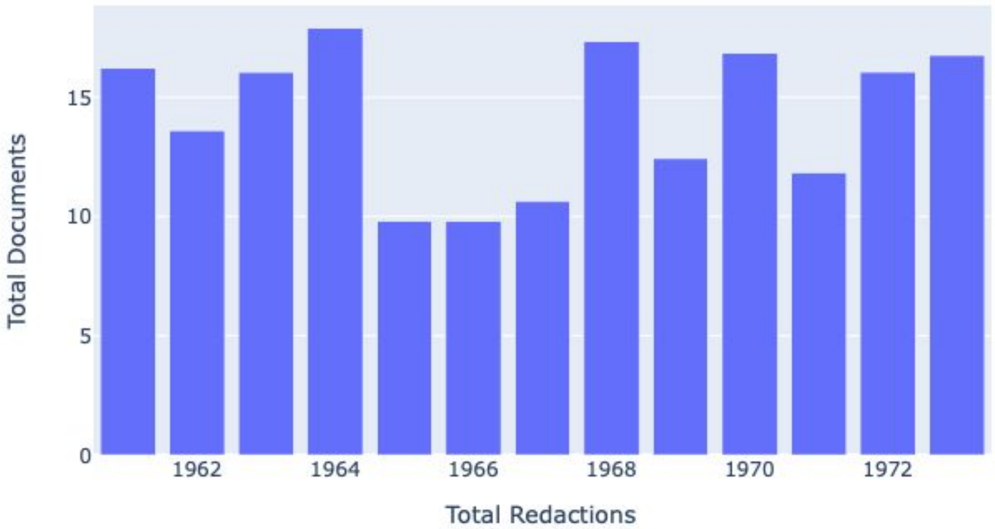
Vietnam


Total Documents
0
5
10
15
1962
1964
1966
1968
1970
1972
Total Redactions

Topics - Weighted Redactions per Document



Moscow


15
10
5
0
Total Documents
1962
1964
1966
1968
1970
1972
Total Redactions