Instagram Policy

```
[0.1] import pdfplumber
[0.2] import re, sys
[0.3] from collections import Counter
[0.4] from difflib import SequenceMatcher

[0.5] PDF_PATH = "/mnt/data/Instagram Privacy Policy.pdf"
[0.6] OUT_PATH = "/mnt/data/Instagram_Privacy_Policy_cleaned.txt"

[0.7] print("Opening:", PDF_PATH)
[0.8] with pdfplumber.open(PDF_PATH) as pdf:
[0.9]     pages = [p.extract_text() or "" for p in pdf.pages]
[0.10] print("Extracted", len(pdf.pages), "pages")

[0.11] def normalize(text):
[0.12]     return [ln.rstrip() for ln in text.replace("\r\n", "\n").replace("\r", "\n").split("\n")]

[0.13] def strip_header_footer(pages, min_frac=0.3):
[0.14]     n = len(pages)
[0.15]     if n < 3:
[0.16]         return pages
[0.17]     from collections import Counter
[0.18]     top_cnt = Counter()
[0.19]     bot_cnt = Counter()
[0.20]     for p in pages:
[0.21]         lines = normalize(p)
[0.22]         for ln in lines[:3]:
[0.23]             if len(ln) > 5:
[0.24]                 top_cnt[ln] += 1
[0.25]         for ln in lines[-3:]:
[0.26]             if len(ln) > 5:
[0.27]                 bot_cnt[ln] += 1
[0.28]     thr = max(2, int(min_frac * n))
[0.29]     top_drop = {ln for ln, c in top_cnt.items() if c >= thr}
[0.30]     bot_drop = {ln for ln, c in bot_cnt.items() if c >= thr}
[0.31]     print("Header lines dropped:", len(top_drop))
[0.32]     print("Footer lines dropped:", len(bot_drop))
[0.33]     for ln in sort
```

```
[0.34] sorted(top_drop, key=top_cnt.get, reverse=True)[:5]:
[0.35]     print("  H:", repr(ln))
[0.36] for ln in sorted(bot_drop, key=bot_cnt.get, reverse=True)[:5]:
[0.37]     print("  F:", repr(ln))
[0.38]     return [[ln for ln in normalize(p) if ln not in top_drop and ln not in bot_drop] for p in pages]

[0.39] cleaned = strip_header_footer(pages)

[0.40] def collapse(lines):
[0.41]     out, prev_blank = [], False
[0.42]     for ln in lines:
[0.43]         blank = (ln.strip == "")
[0.44]         if blank and prev_blank:
[0.45]             continue
[0.46]         out.append(ln)
[0.47]         prev_blank = blank
[0.48]     return out

[0.49] cleaned = [collapse(p) for p in cleaned]

[0.50] raw = "\n\n".join("\n".join(p) for p in cleaned)

[0.51] print("Saving:", OUT_PATH)
[0.52] open(OUT_PATH, "w", encoding="utf-8").write(raw)

[0.53] print("Total chars:", len(raw))
[0.54] print("Total lines:", raw.count("\n"))

[0.55] print("\n--- preview (first 2000 chars) ---")
[0.56] print(raw[:2000])
```

```
Opening: /mnt/data/Instagram Privacy Policy.pdf
Extracted 14 pages
Header lines dropped: 0
Footer lines dropped: 0
  F: 'Instagram Privacy Policy'
  F: '2'
Saving: /mnt/data/Instagram_Privacy_Policy_cleaned.txt
Total chars: 35777
Total lines: 1 --- preview (first 2000 chars) ---
Instagram Privacy Policy

This policy describes the information we process
to support Facebook products, services and
integrations. To learn more about how we process
information in connection with AI at Meta, visit
the AI Info page.

Contents

Introduction
What kinds of information do we collect?
How do we use this information?
How do we share this information?
How do we transfer data across borders?
How do we collect this information?
How do we use cookies and similar technologies?
How do we respond to legal requests or prevent harm?
How long do we keep your information?
How do we prevent minors from using Instagram?
How do we notify you of changes to this policy?
How to contact Meta with questions
Additional Notice for California residents
Additional Notice for Jurisdictions in the European Region
Additional Notice for Jurisdictions in the United Kingdom
Additional Notice for Jurisdictions in the Rest of Europe
Additional Notice for Jurisdictions in Brazil
Additional Notice for Jurisdictions in Thailand
Additional Notice for Jurisdictions in the United States
Additional Notice for Jurisdictions in Japan
Additional Notice for Jurisdictions in Australia
Additional Notice for Jurisdictions in Korea
Additional Notice for Jurisdictions in Canada
Additional Notice for Jurisdictions in Argentina
Additional Notice for Jurisdictions in Chile
Additional Notice for Jurisdictions in Colombia
Additional Notice for Jurisdictions in Indonesia
Additional Notice for Jurisdictions in Malaysia
Additional Notice for Jurisdictions in Mexico
Additional Notice for Jurisdictions in New Zealand
Additional Notice for Jurisdictions in Philippines
Additional Notice for Jurisdictions in Singapore
Additional Notice for Jurisdictions in Vietnam
Additional Notice for Jurisdictions in Saudi Arabia
Additional Notice for Jurisdictions in United Arab Emirates
Additional Notice for Jurisdictions in Bahrain
Additional Notice for Jurisdictions in Jordan
Additional Notice for Jurisdictions in Kuwait
Additional Notice for Jurisdictions in Oman
Additional Notice for Jurisdictions in Qatar
Additional Notice for Jurisdictions in Algeria
Additional Notice for Jurisdictions in Egypt
Additional Notice for Jurisdictions in Iraq
Additional Notice for Jurisdictions in Morocco
Additional Notice for Jurisdictions in Tunisia
Additional Notice for Jurisdictions in Yemen
Additional Notice for Jurisdictions in Israel
Additional Notice for Jurisdictions in Kazakhstan
Additional Notice for Jurisdictions in Pakistan
Additional Notice for Jurisdictions in Uzbekistan
Additional Notice for Jurisdictions in Cambodia
Additional Notice for Jurisdictions in Hong Kong
Additional Notice for Jurisdictions in Macau
Additional Notice for Jurisdictions in Mongolia
Additional Notice for Jurisdictions in Taiwan
Additional Notice for Jurisdictions in Bangladesh
Additional Notice for Jurisdictions in India
Additional Notice for Jurisdictions in Nepal
Additional Notice for Jurisdictions in Sri Lanka
Additional Notice for Jurisdictions in Australia, New Zealand and Pacific Islands
Additional Notice for Jurisdictions in Bolivia
Additional Notice for Jurisdictions in Costa Rica
Additional Notice for Jurisdictions in Dominican Republic
Additional Notice for Jurisdictions in Ecuador
Additional Notice for Jurisdictions in El Salvador
Additional Notice for Jurisdictions in Guatemala
Additional Notice for Jurisdictions in Honduras
Additional Notice for Jurisdictions in Nicaragua
Additional Notice for Jurisdictions in Paraguay
Additional Notice for Jurisdictions in Peru
Additional Notice for Jurisdictions in Uruguay
Additional Notice for Jurisdictions in Venezuela
Additional Notice for Jurisdictions in Brunei
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor
incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud
exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure
dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.
Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt
mollit anim id est laborum.
```

The fourth page is a sample of the effect of sentence, the square sentence with full stop. They are more than one full stop and the sentence is longer is the end. Here we are repeating such sentence for better pattern matching.

In [16]:
```
lineno = 1
```

In [17]:
```
def myprint(text):
    global lineno
    print lineno,
    lineno = lineno + 1
    print text,
    sys.stdout.flush()
    return sys.stdout.write.__doc__[0:0]
```

In [18]:
```
lineno = 1
f = open('/tmp/output')
f.read(1000)
```

Executing the python matched in the line

In [19]:
```
print f.read(1000)
```

```
1 Lorem ipsum dolor
2 sit amet, consectetur
3 adipiscing elit, sed do eiusmod
4 tempor incididunt ut
5 labore et dolore magna
6 aliqua. Ut enim ad minim
7 veniam, quis nostrud
8 exercitation ullamco laboris
9 nisi ut aliquip ex ea
10 commodo consequat. Duis aute
11 irure dolor in reprehenderit
12 in voluptate velit esse
13 cillum dolore eu fugiat nulla
14 pariatur. Excepteur sint
15 occaecat cupidatat non proident,
16 sunt in culpa qui officia
17 deserunt mollit anim id
18 est laborum.
19 Lorem ipsum dolor sit amet,
20 consectetur adipiscing elit,
21 sed do eiusmod tempor
22 incididunt ut labore et
23 dolore magna aliqua. Ut enim
24 ad minim veniam, quis nostrud
25 exercitation ullamco laboris
26 nisi ut aliquip ex ea
27 commodo consequat. Duis aute
28 irure dolor in reprehenderit
29 in voluptate velit esse
30 cillum dolore eu fugiat nulla
31 pariatur. Excepteur sint
32 occaecat cupidatat non proident,
33 sunt in culpa qui officia
34 deserunt mollit anim id
35 est laborum.
36 Lorem ipsum dolor sit amet,
37 consectetur adipiscing elit,
38 sed do eiusmod tempor
39 incididunt ut labore et
40 dolore magna aliqua. Ut enim
41 ad minim veniam, quis nostrud
42 exercitation ullamco laboris
43 nisi ut aliquip ex ea
44 commodo consequat. Duis aute
45 irure dolor in reprehenderit
46 in voluptate velit esse
47 cillum dolore eu fugiat nulla
48 pariatur. Excepteur sint
49 occaecat cupidatat non proident,
50 sunt in culpa qui officia
51 deserunt mollit anim id
52 est laborum.
53 Lorem ipsum dolor sit amet,
54 consectetur adipiscing elit,
55 sed do eiusmod tempor
56 incididunt ut labore et
57 dolore magna aliqua. Ut enim
58 ad minim veniam, quis nostrud
59 exercitation ullamco laboris
60 nisi ut aliquip ex ea
61 commodo consequat. Duis aute
62 irure dolor in reprehenderit
63 in voluptate velit esse
64 cillum dolore eu fugiat nulla
65 pariatur. Excepteur sint
66 occaecat cupidatat non proident,
67 sunt in culpa qui officia
68 deserunt mollit anim id
69 est laborum.
```

The sentence ends after the mark. The first cell shows the result in the myprint cell and the second cell is just the dataframe of the result.

In [20]:
```
lineno = 1
f.close()
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
```

In [21]:
```
f = open('/tmp/output')
f.read(1000)
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
```

In [22]:
```
f.close()
```

In [23]:
```
lineno = 1
f = open('/tmp/output')
```

In [24]:
```
print f.read(1000)
```

```
1 Lorem ipsum dolor
2 sit amet, consectetur
3 adipiscing elit, sed do eiusmod
4 tempor incididunt ut
5 labore et dolore magna
6 aliqua. Ut enim ad minim
7 veniam, quis nostrud
8 exercitation ullamco laboris
9 nisi ut aliquip ex ea
10 commodo consequat. Duis aute
11 irure dolor in reprehenderit
12 in voluptate velit esse
13 cillum dolore eu fugiat nulla
14 pariatur. Excepteur sint
15 occaecat cupidatat non proident,
16 sunt in culpa qui officia
17 deserunt mollit anim id
18 est laborum.
19 Lorem ipsum dolor sit amet,
20 consectetur adipiscing elit,
21 sed do eiusmod tempor
22 incididunt ut labore et
23 dolore magna aliqua. Ut enim
24 ad minim veniam, quis nostrud
25 exercitation ullamco laboris
26 nisi ut aliquip ex ea
27 commodo consequat. Duis aute
28 irure dolor in reprehenderit
29 in voluptate velit esse
30 cillum dolore eu fugiat nulla
31 pariatur. Excepteur sint
32 occaecat cupidatat non proident,
33 sunt in culpa qui officia
34 deserunt mollit anim id
35 est laborum.
36 Lorem ipsum dolor sit amet,
37 consectetur adipiscing elit,
38 sed do eiusmod tempor
39 incididunt ut labore et
40 dolore magna aliqua. Ut enim
41 ad minim veniam, quis nostrud
42 exercitation ullamco laboris
43 nisi ut aliquip ex ea
44 commodo consequat. Duis aute
45 irure dolor in reprehenderit
46 in voluptate velit esse
47 cillum dolore eu fugiat nulla
48 pariatur. Excepteur sint
49 occaecat cupidatat non proident,
50 sunt in culpa qui officia
51 deserunt mollit anim id
52 est laborum.
53 Lorem ipsum dolor sit amet,
54 consectetur adipiscing elit,
55 sed do eiusmod tempor
56 incididunt ut labore et
57 dolore magna aliqua. Ut enim
58 ad minim veniam, quis nostrud
59 exercitation ullamco laboris
60 nisi ut aliquip ex ea
61 commodo consequat. Duis aute
62 irure dolor in reprehenderit
63 in voluptate velit esse
64 cillum dolore eu fugiat nulla
65 pariatur. Excepteur sint
66 occaecat cupidatat non proident,
67 sunt in culpa qui officia
68 deserunt mollit anim id
69 est laborum.
```

In [25]:
```
f.close()
```

```
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
```

In [26]:
```
df = pd.read_csv('/tmp/output', sep=',', header=None)
df.head()
```

	0
0	Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
1	Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.
2	Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.