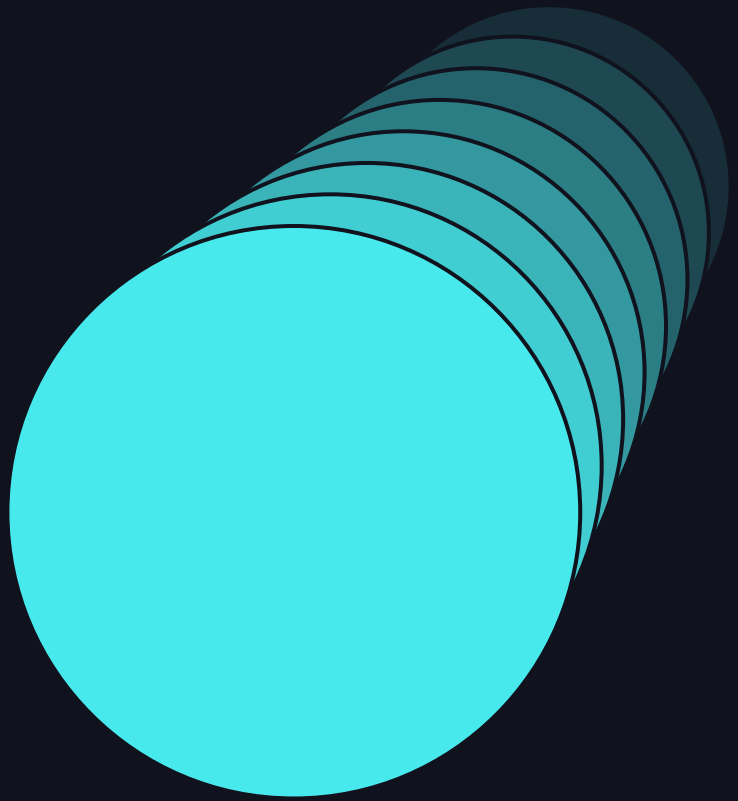
DATA+AI SUMMIT BY databricks

Best Practices for Data Prep for GenAI Development



Brian Kihoon Lee

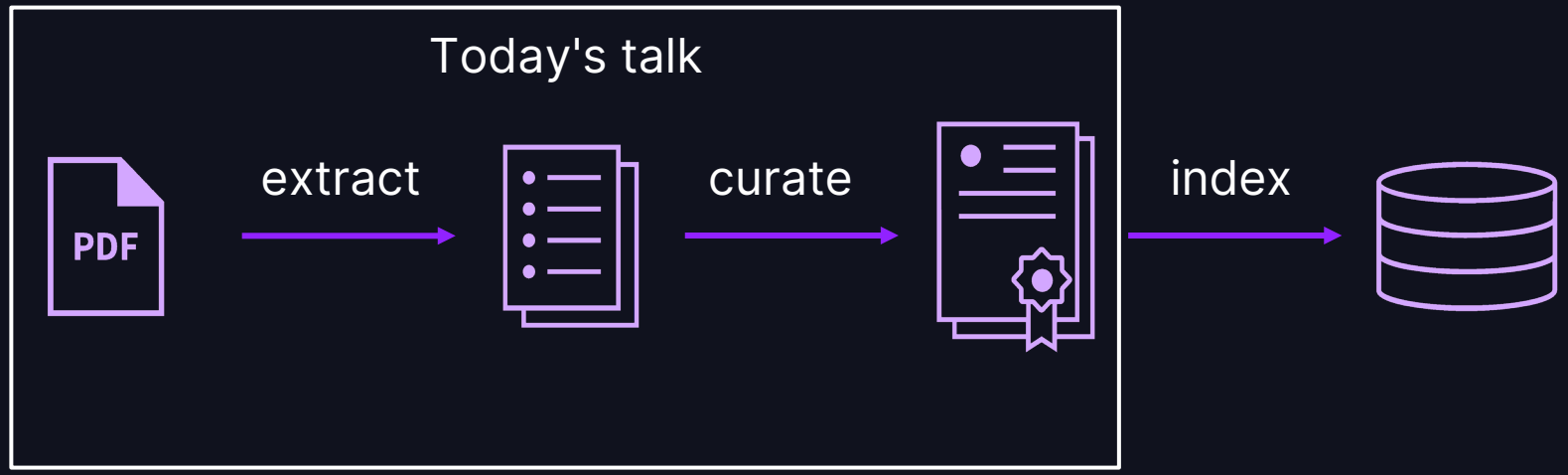
©2024 Databricks Inc. — All rights reserved

Lilac x Databricks

- Lilac joined Databricks in March 2024
- We built LLM-superpowered tools to understand and curate text data

©2024 Databricks Inc. — All rights reserved

A typical RAG data pipeline



©2024 Databricks Inc. — All rights reserved

LIVE DEMO: EXTRACTION

©2024 Databricks Inc. — All rights reserved

Extract data from PDF

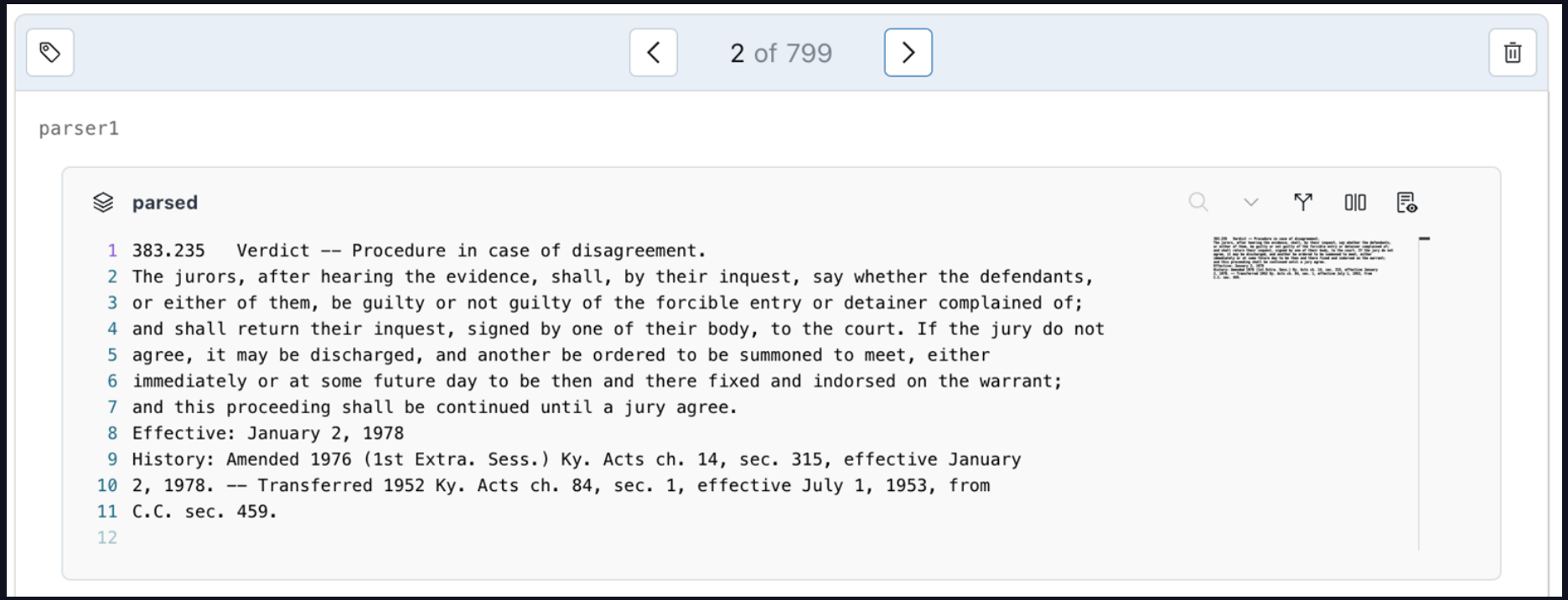
PYTHON

```
def parsed1(filepath):
  try:
    doc = pymupdf.open(filepath)  # open a document
    output = '\n'.join(page.get_text() for page in doc)
  except Exception as e:
    print(e, filepath)
    output = ''
  return {'parsed': output}

dataset.map(
  parsed1, input_path='filepath', output_path='parser1', num_jobs=8, execution_type='processes', overwrite=True
)
```


©2024 Databricks Inc. — All rights reserved

Look at your data



©2024 Databricks Inc. — All rights reserved

Extract data from PDF, attempt #2

PYTHON

```
def parsed2(filepath):
  try:
    doc = pymupdf.open(filepath)  # open a document
    text_builder = []
    for page in doc:
      blocks = page.get_text('blocks')
      text_builder.append('\n\n'.join(block[4].replace('\n', ' ') for block in blocks))
    output = ''.join(text_builder)
  except Exception as e:
    print(e, filepath)
    output = ''
  return {'parsed': output}

dataset.map(parsed2, input_path='filepath', output_path='parser2', num_jobs=8, execution_type='processes')
```


©2024 Databricks Inc. — All rights reserved

Don't overfit to one document

parser1

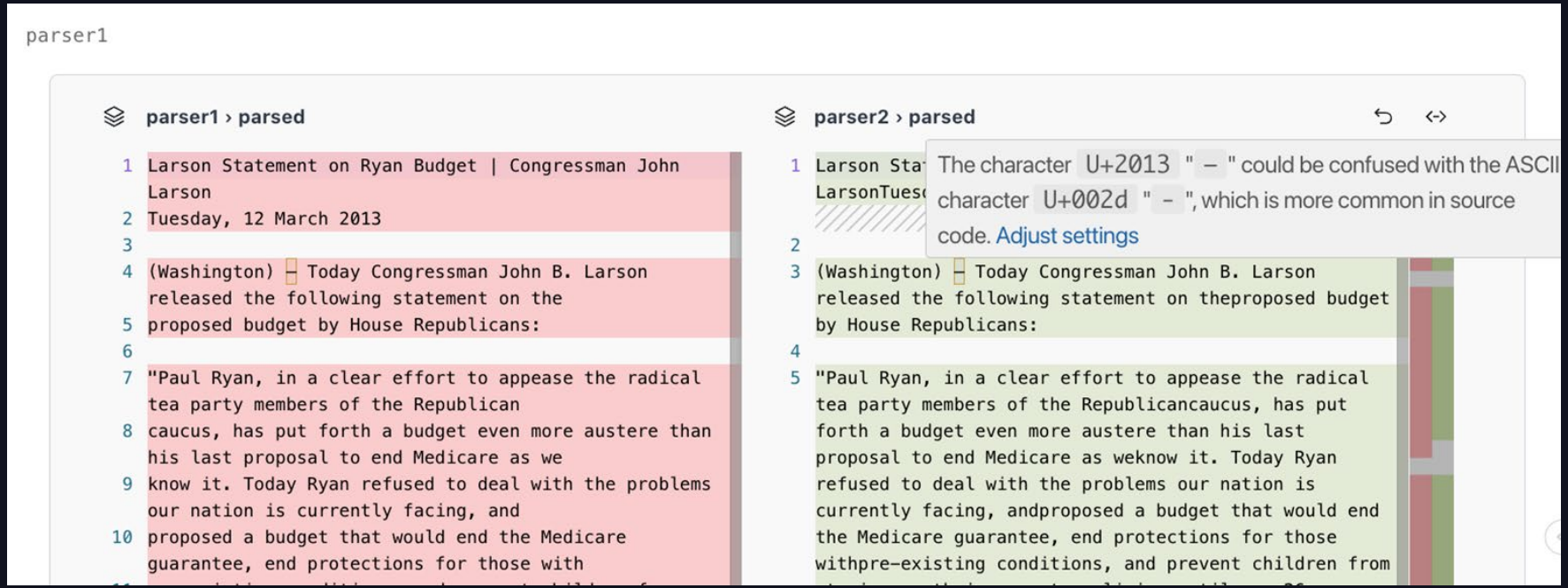
parser1 › parsed

```
1  383.235   Verdict -- Procedure in case of
   disagreement.
2  The jurors, after hearing the evidence, shall, by
   their inquest, say whether the defendants,
3  or either of them, be guilty or not guilty of the
   forcible entry or detainer complained of;
4  and shall return their inquest, signed by one of their
   body, to the court. If the jury do not
5  agree, it may be discharged, and another be ordered to
   be summoned to meet, either
6  immediately or at some future day to be then and there
   fixed and indorsed on the warrant;
7  and this proceeding shall be continued until a jury
   agree.
8  Effective: January 2, 1978
9  History: Amended 1976 (1st Extra. Sess.) Ky. Acts ch.
   14, sec. 315, effective January
10 2, 1978. -- Transferred 1952 Ky. Acts ch. 84, sec. 1,
   effective July 1, 1953, from
11 C.C. sec. 459.
12
```

parser2 › parsed

```
1  383.235   Verdict -- Procedure in case of
   disagreement.
2  The jurors, after hearing the evidence, shall, by
   their inquest, say whether the defendants, or either
   of them, be guilty or not guilty of the forcible entry
   or detainer complained of; and shall return their
   inquest, signed by one of their body, to the court. If
   the jury do not agree, it may be discharged, and
   another be ordered to be summoned to meet, either
   immediately or at some future day to be then and there
   fixed and indorsed on the warrant; and this proceeding
   shall be continued until a jury agree.
3  Effective: January 2, 1978
4  History: Amended 1976 (1st Extra. Sess.) Ky. Acts ch.
   14, sec. 315, effective January
5  2, 1978. -- Transferred 1952 Ky. Acts ch. 84, sec. 1,
   effective July 1, 1953, from C.C. sec. 459.
```

Normalize whitespace and unicode



©2024 Databricks Inc. — All rights reserved

Clean up data from PDF

PYTHON

```
from unstructured.cleaners.core import clean_bullets, replace_unicode_quotes
import re

def clean_text(text):
  text = replace_unicode_quotes(text)  # Homogenize quotation marks
  text = clean_bullets(text)  # Homogenize bullet points
  text = re.sub(r'[\u00ad]', '', text)  # Soft hyphen removal
  text = re.sub(r'[\u2013\u2212]', '-', text)  # Uncommon hyphen variants
  text = re.sub(r'[ ]{2,}', ' ', text)  # Normalize spacing
  text = re.sub(r'[\u00a0]', ' ', text)  # Remove invisible whitespace
  return text

dataset.map(clean_text, input_path='parser2.parsed', output_path='parser2.cleaned', overwrite=True)
```

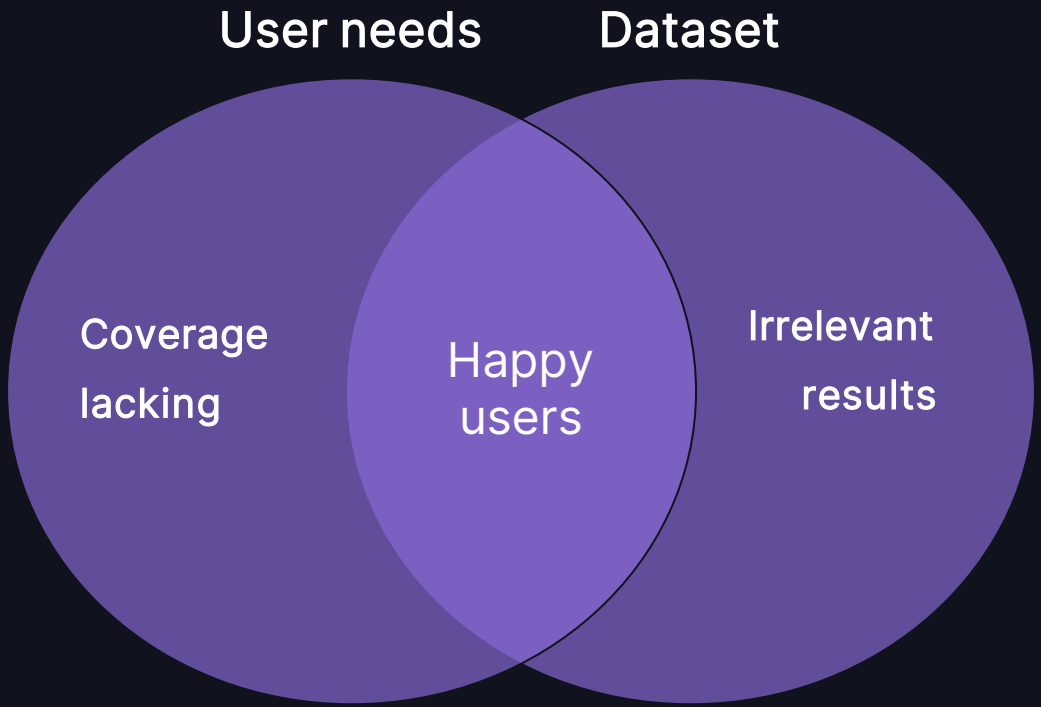

©2024 Databricks Inc. — All rights reserved

Extraction: summary

1. Look at your data!
2. Don't overfit your pipeline to a few documents
3. Normalize whitespace and uncommon characters

©2024 Databricks Inc. — All rights reserved

What is curation?

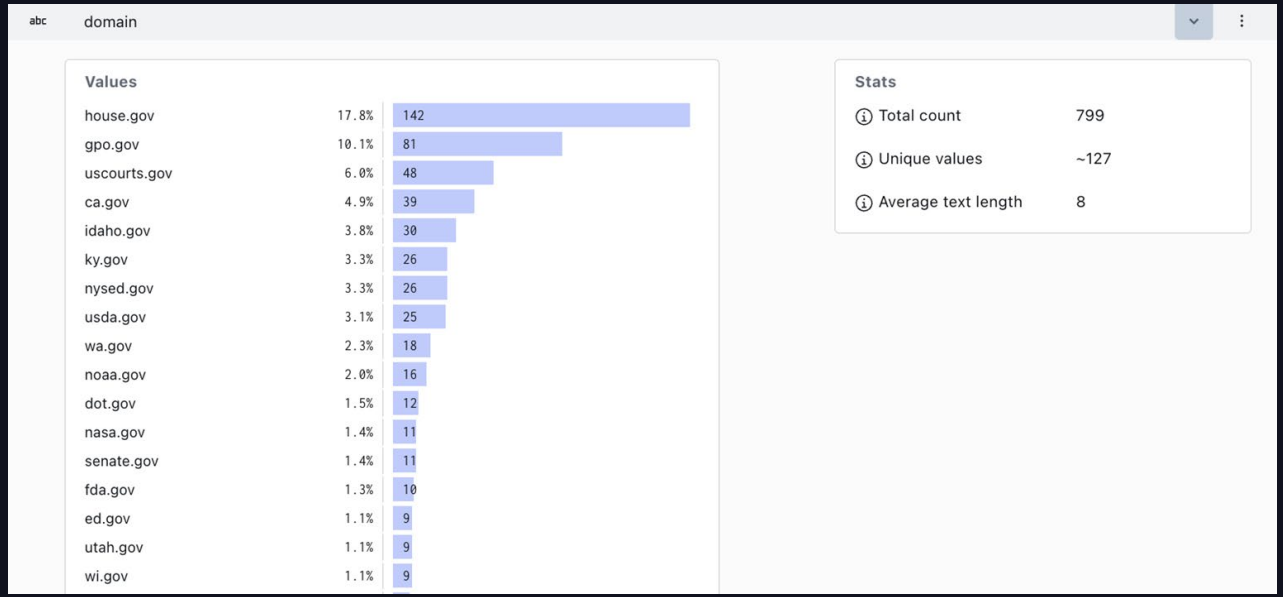


©2024 Databricks Inc. — All rights reserved

LIVE DEMO: CURATION

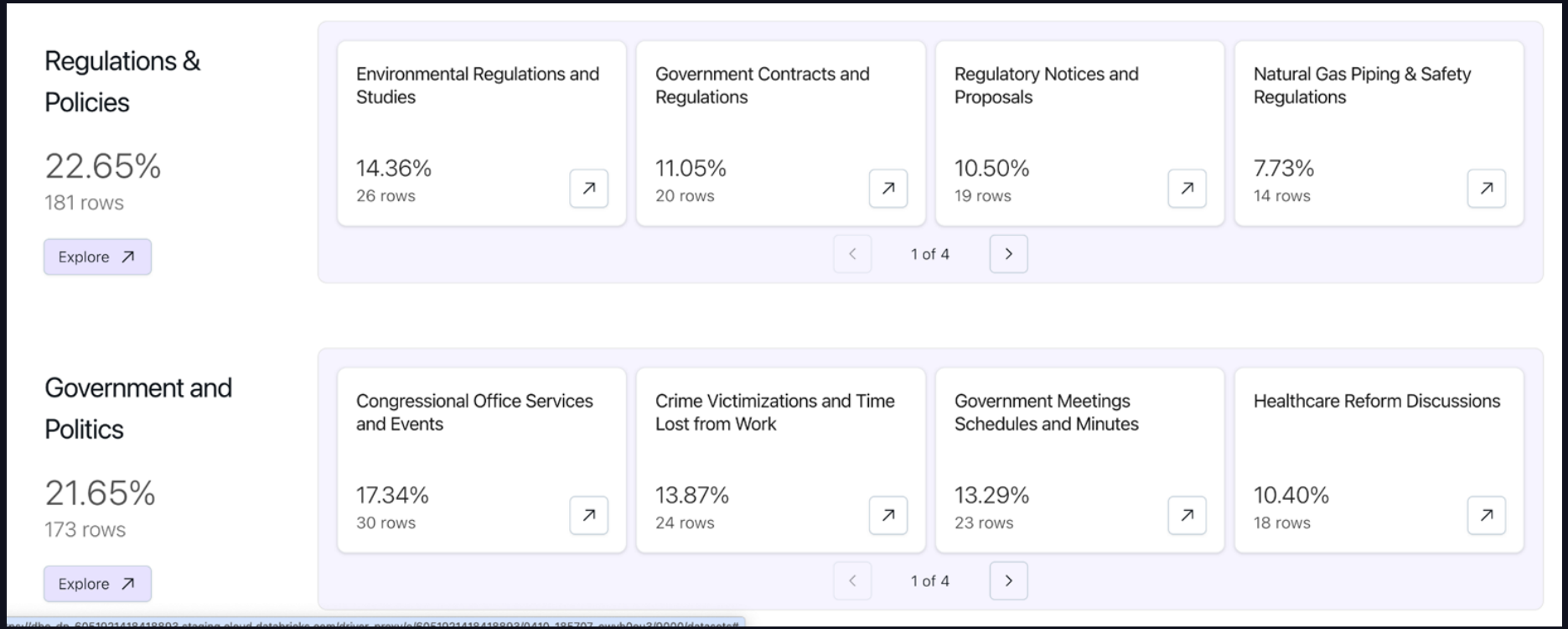
©2024 Databricks Inc. — All rights reserved

Start with structured columns



©2024 Databricks Inc. — All rights reserved

Clustering creates structure



©2024 Databricks Inc. — All rights reserved

Curation: summary

1. Every dataset and use case is different: you are the domain expert
2. Use histograms and text clusters to check coverage and relevancy
3. Cross-reference usage logs with your indexed data

©2024 Databricks Inc. — All rights reserved

Takeaways

Tools used in this talk:

PDF extraction: PyMuPDF, unstructured
Text cleaning: spacy, unstructured, regex
Visualization: Lilac

Extraction

1. Look at your data!
2. Don't overfit your pipeline to a few documents
3. Normalize whitespace and uncommon characters

Curation

1. Every dataset is different: you are the domain expert
2. Use histograms and text clusters to check coverage and relevancy
3. Cross-reference usage logs with your indexed data

©2024 Databricks Inc. — All rights reserved