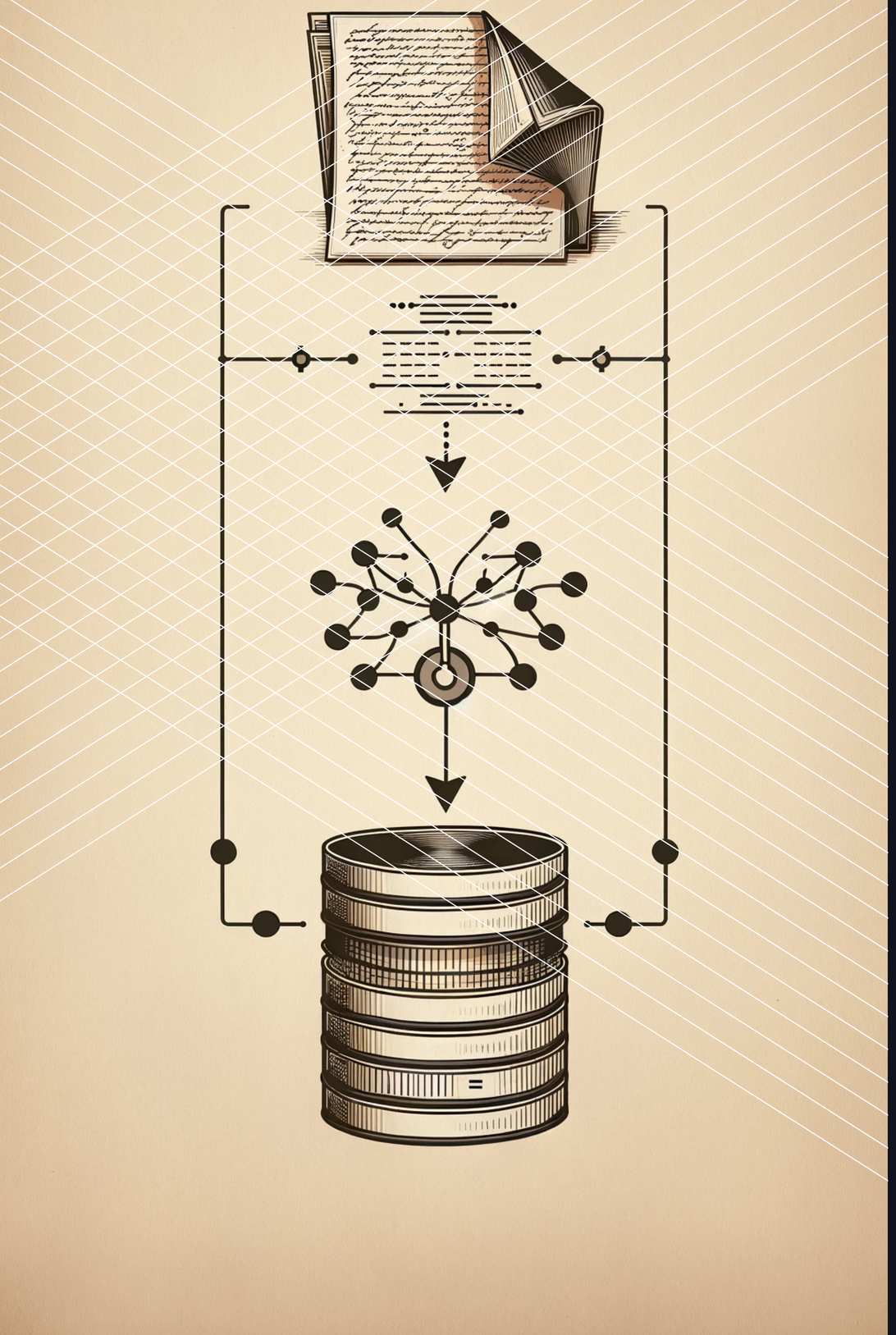


November 2023

tS tranSkriptorium

Beyond automatic transcription:

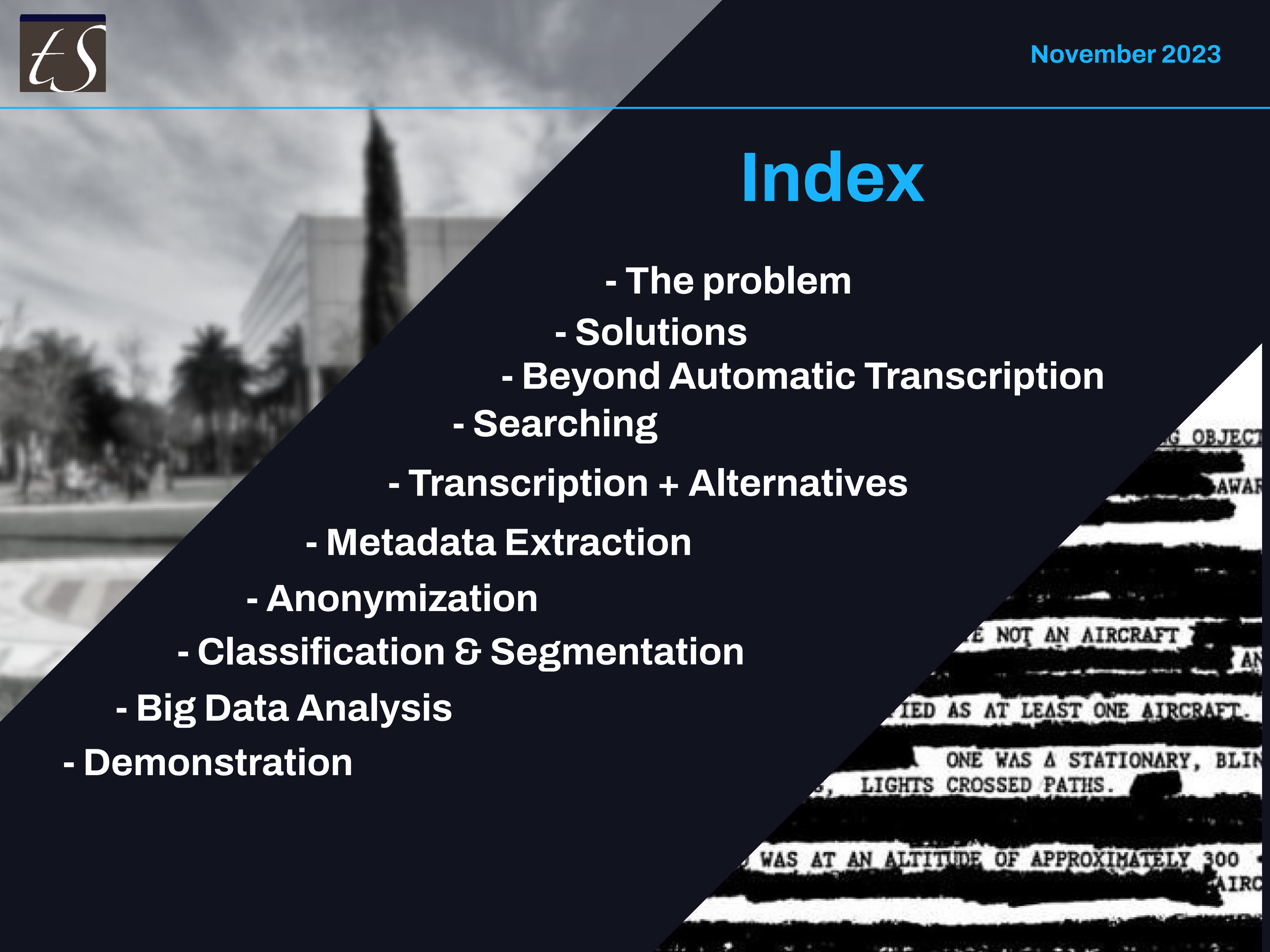
A better source of information on which to apply NLP with success

www.transkriptorium.com

enisa
CERTIFICADA COMO EMPRESA EMERGENTE
MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO

November 2023

Index



November 2023

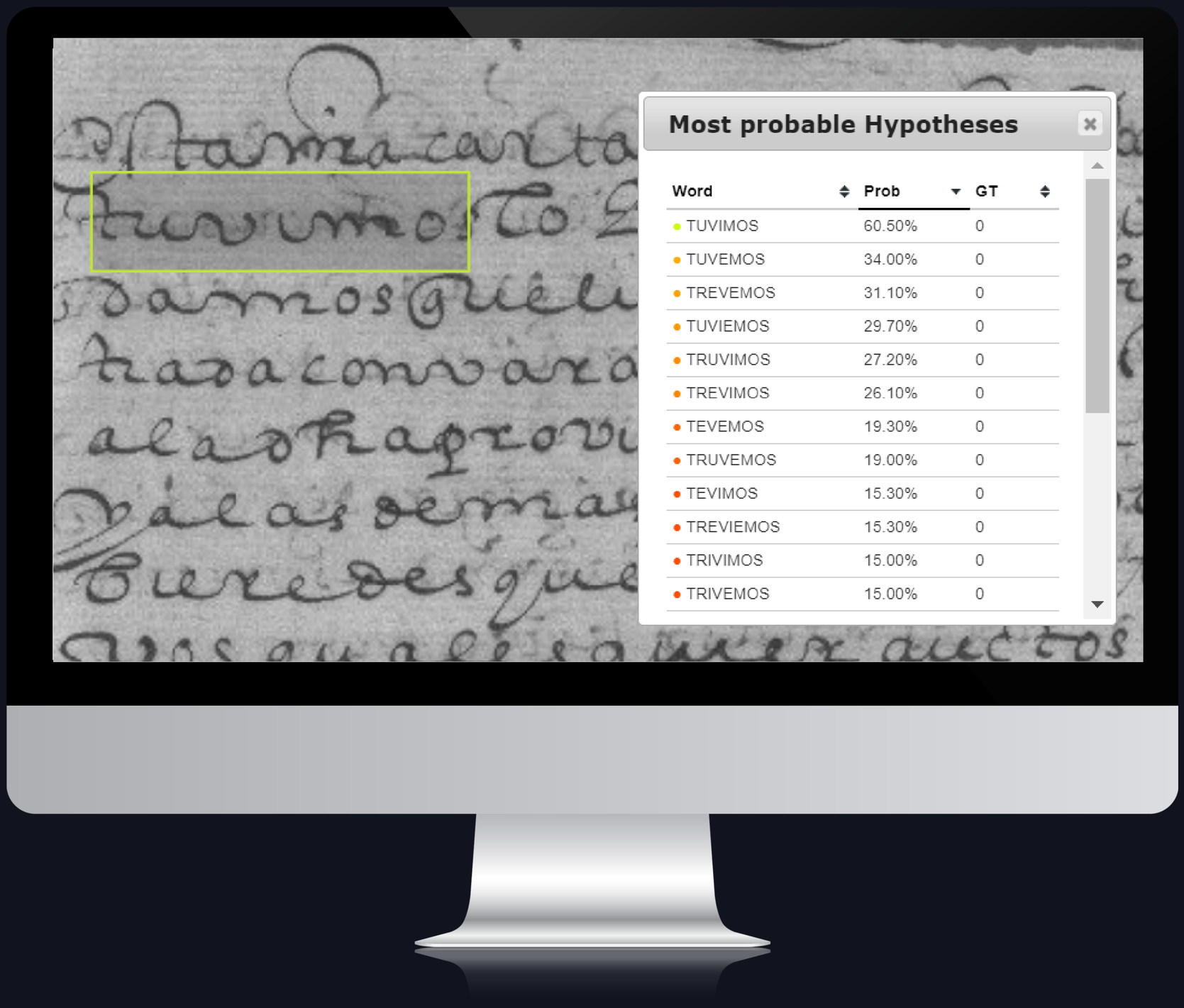
The Problem

- Large collections of physical documents
- Some are still growing
- Handwritten, Typed, printed and mixed
- Structured, Semi-structured, Unstructured
- Changing Templates
- Different access needs depending on the collection
- Depending on the country some laws are in place

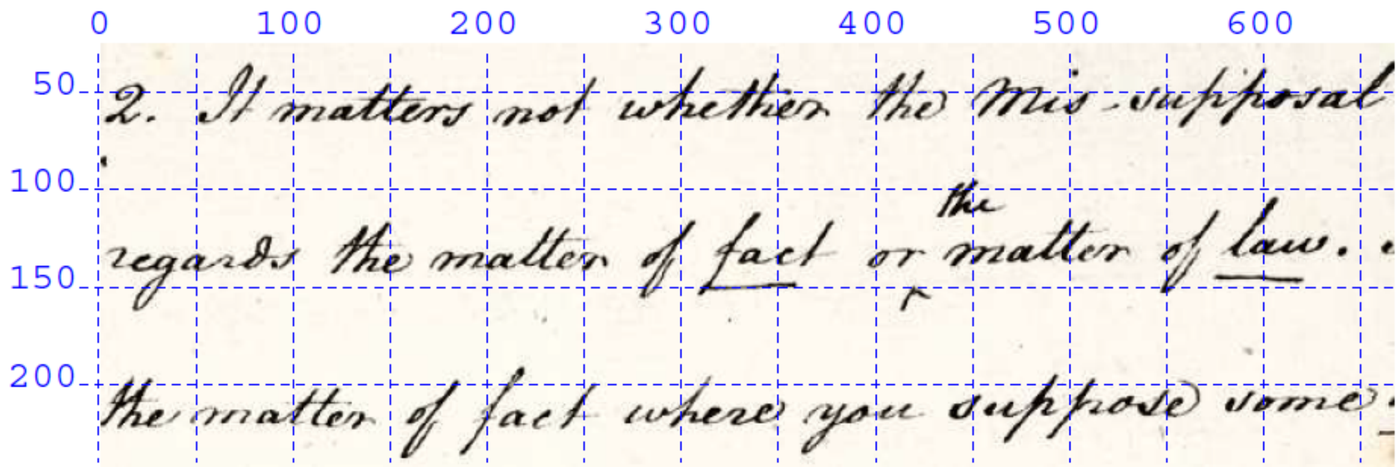
Solutions

- Manual Process
- Automatic Transcription
- You may not need it
- Hybrid Approach

Beyond Automatic Transcription



Probabilistic Indexing

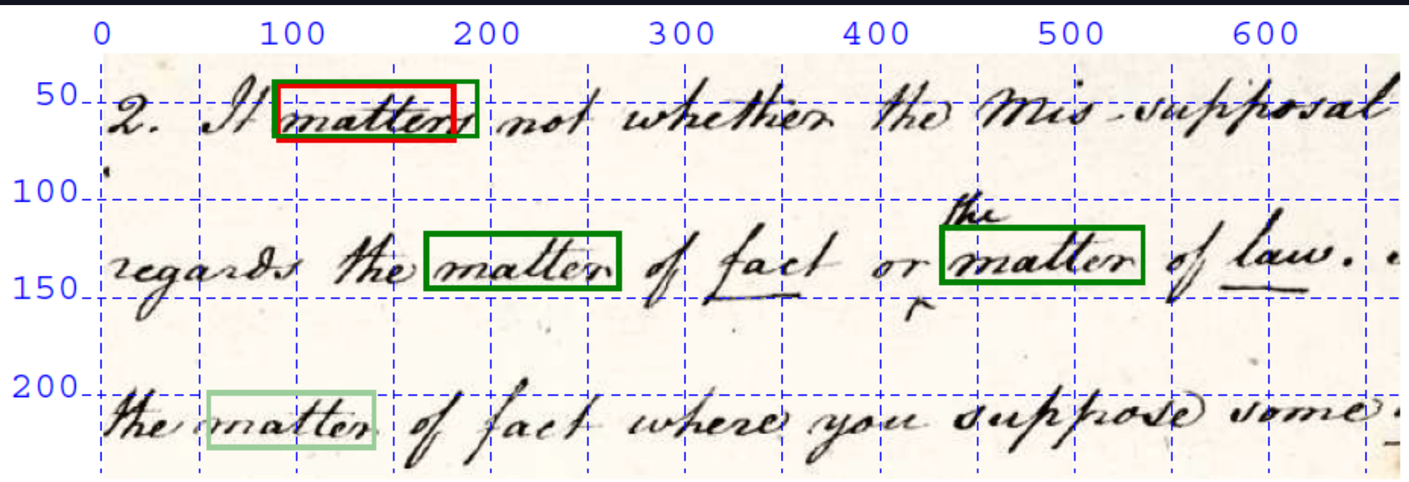



```
# pageID="Bentham-071-021-002-part"
#    keyword relPrb    bounding box
#
          2  0.929    1  36  20 31
         21  0.064    1  36  24 31
         IT  0.982   33  36  27 31
         IF  0.012   33  36  26 31
    MATTERS  0.989   76  35 104 31
     MATTER  0.011   77  36  93 31
        NOT  0.999  216  36  47 31
    WHETHER  1.000  256  36  99 31
        THE  0.997  389  36  33 31
MIS-SUPPOSAL 1.000  455  36 193 31

        THE  0.927  430  88  30 31
        LHE  0.056  434  88  25 31
        ...  ...      ...     ...
```

```
REGARDS  0.857    5 115  84 31
UGARDS   0.138    5 115  80 31
    THE  0.993  110 115  43 31
 MATTER  0.998  160 115  93 31
     OF  0.996  271 115  23 31
   FACT  0.999  306 115  49 31
     OR  0.973  377 115  37 31
     ON  0.021  377 115  42 31
 MATTER  0.990  425 116 100 31
     OF  0.995  542 115  25 31
    LAM  0.407  575 115  30 31
   BIMR  0.175  575 115  55 31
    ...  ...      ...     ...
    LAW  0.032  575 115  36 31
   TAUE  0.031  575 115  55 31
    ...  ...      ...     ...
   LANE  0.012  575 115  59 31
```

```
    THE  0.990    1 198  28 31
 MATTER  0.934   61 198  64 31
     OF  0.988  141 198  28 31
   FAST  0.367  182 198  62 31
    FAR  0.186  182 198  36 31
    ...  ...      ...     ...
   FACT  0.017  182 198  46 31
     AS  0.142  200 198  29 31
    HAE  0.022  200 198  29 31
  WHERE  0.992  255 198  90 31
    YOU  0.761  365 198  45 31
    YOW  0.030  365 198  45 31
   GOUS  0.064  372 198  47 31
SUPPOSE  0.975  429 198 120 31
SUPFROSE 0.024  429 198 125 31
   SOME  0.834  570 198  78 31
  SONER  0.016  576 198  83 31
    OME  0.109  580 198  65 31
     ME  0.022  620 198  22 31
```

Probabilistic Indexing



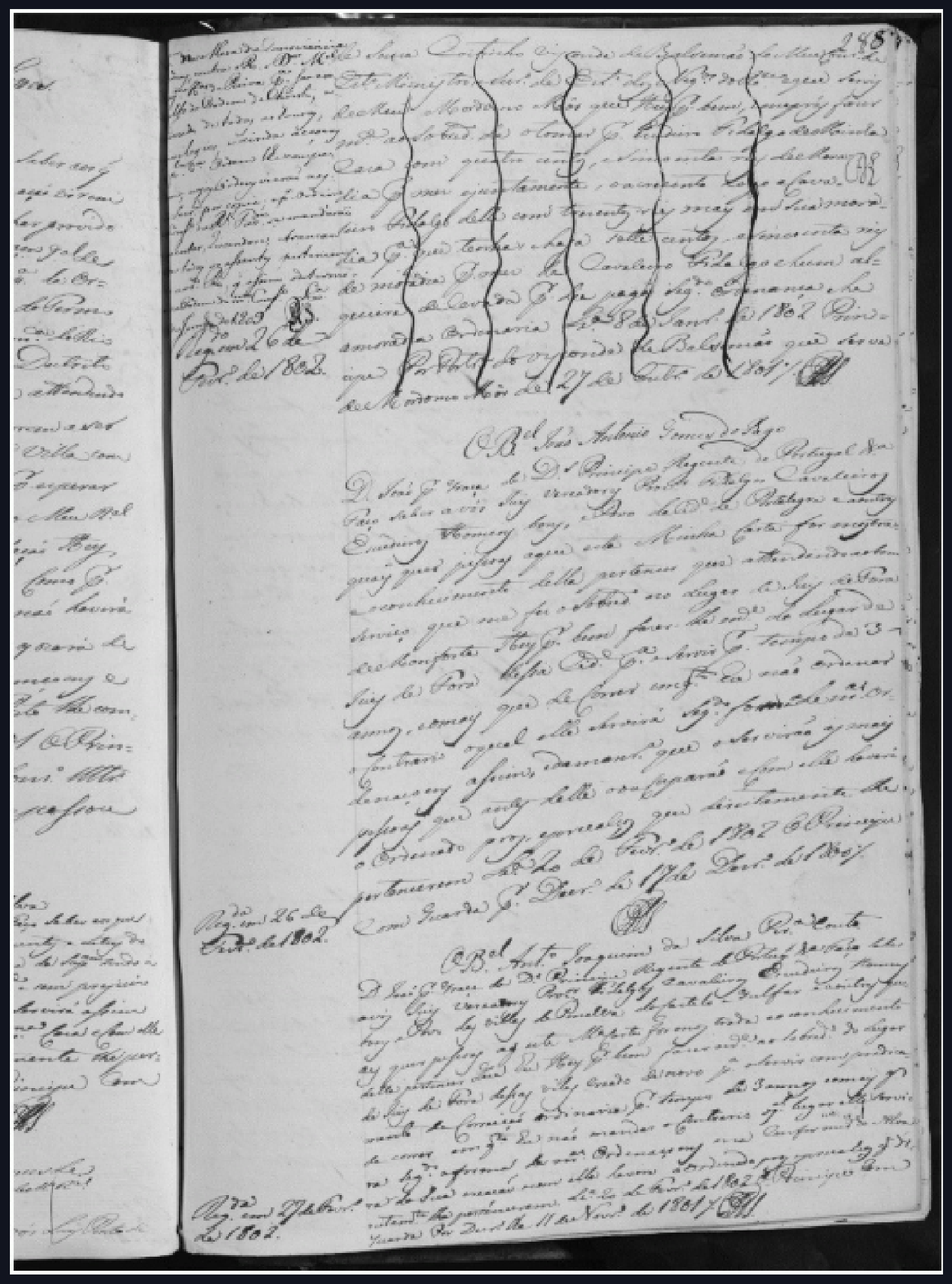

```
# pageID="Bentham-071-021-002-part"
#    keyword relPrb    bounding box
#
          2  0.929    1  36  20 31
         21  0.064    1  36  24 31
         IT  0.982   33  36  27 31
         IF  0.012   33  36  26 31
    MATTERS  0.998   76  35 104 31
     MATTER  0.011   77  36  93 31
        NOT  0.999  216  36  47 31
    WHETHER  1.000  256  36  99 31
        THE  0.997  389  36  33 31
MIS-SUPPOSAL 1.000  455  36 193 31

        THE  0.927  430  88  30 31
        LHE  0.056  434  88  25 31
        ...  ...     ...     ...
```

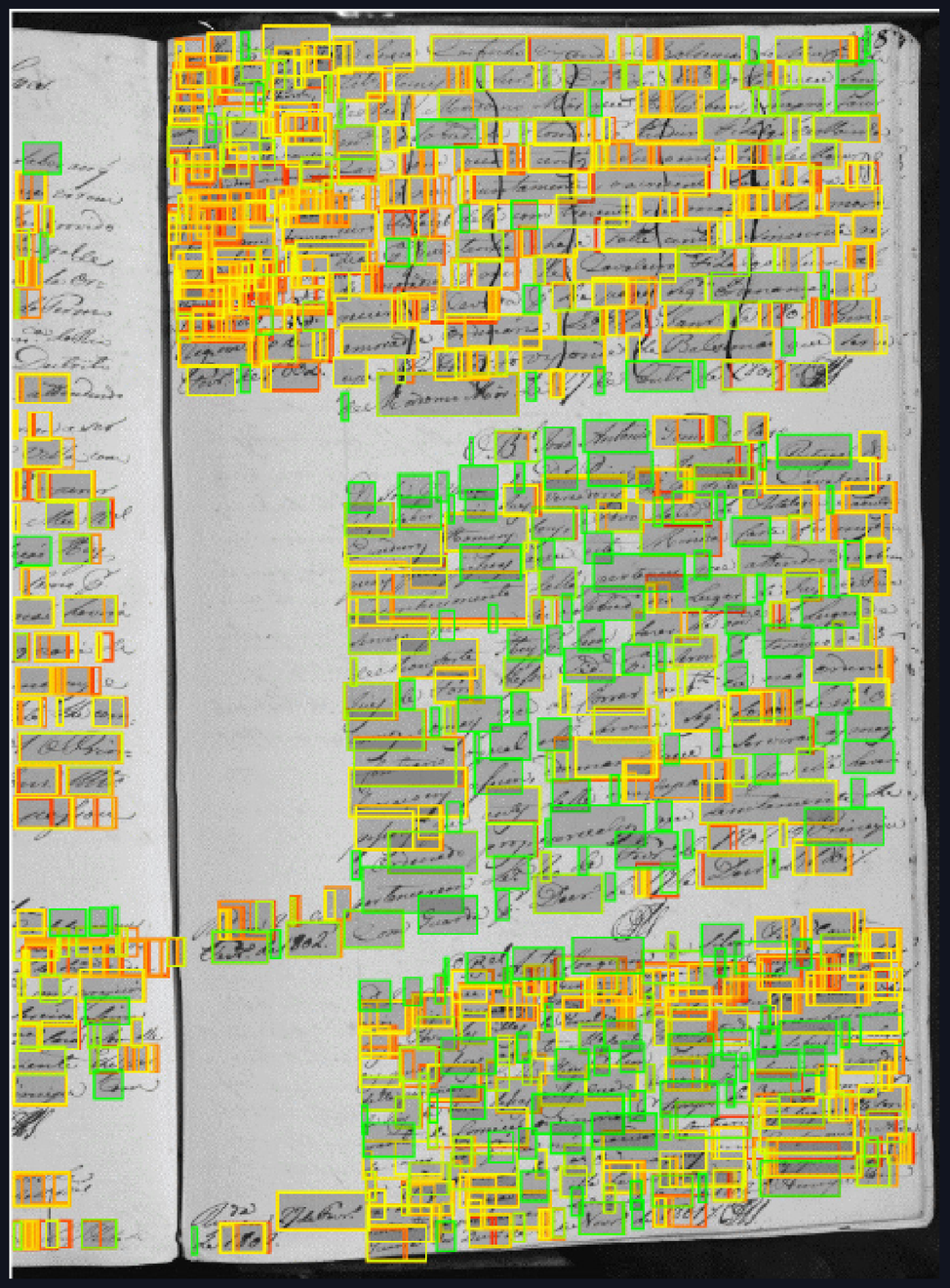
```
REGARDS  0.857    5 115  84 31
UGARDS   0.138    5 115  80 31
    THE  0.993  110 115  43 31
 MATTER  0.998  160 115  93 31
     OF  0.996  271 115  23 31
   FACT  0.999  306 115  49 31
     OR  0.973  377 115  37 31
     ON  0.021  377 115  42 31
 MATTER  0.990  425 116 100 31
     OF  0.995  542 115  25 31
    LAM  0.407  575 115  30 31
   BIMR  0.175  575 115  55 31
    ...  ...      ...     ...
    LAW  0.032  575 115  36 31
   TAUE  0.031  575 115  55 31
    ...  ...      ...     ...
   LANE  0.012  575 115  59 31
```

```
    THE  0.990    1 198  28 31
 MATTER  0.934   61 198  64 31
     OF  0.988  141 198  28 31
   FAST  0.367  182 198  62 31
    FAR  0.186  182 198  36 31
    ...  ...      ...     ...
   FACT  0.017  182 198  46 31
     AS  0.142  200 198  29 31
    HAE  0.022  200 198  29 31
  WHERE  0.992  255 198  90 31
    YOU  0.761  365 198  45 31
    YOW  0.030  365 198  45 31
   GOUS  0.064  372 198  47 31
SUPPOSE  0.975  429 198 120 31
SUPFROSE 0.024  429 198 125 31
   SOME  0.834  570 198  78 31
  SONER  0.016  576 198  83 31
    OME  0.109  580 198  65 31
     ME  0.022  620 198  22 31
```

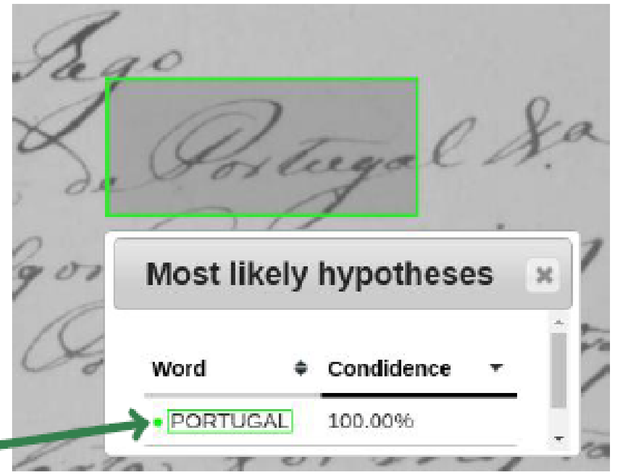
November 2023

Probabilistic Indexing



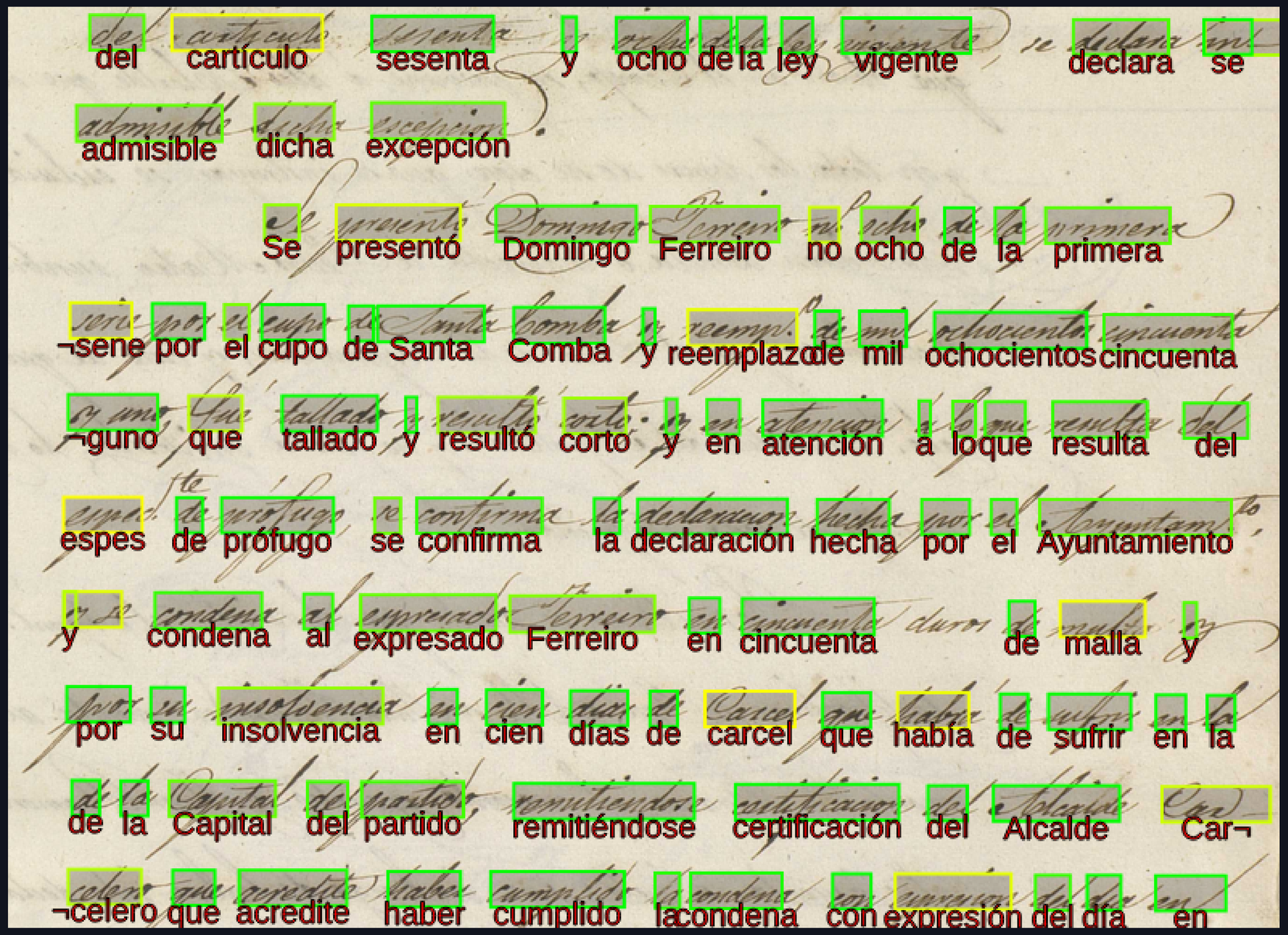


Searching


Most likely hypotheses
Word
Condidence
PORTUGAL
100.00%

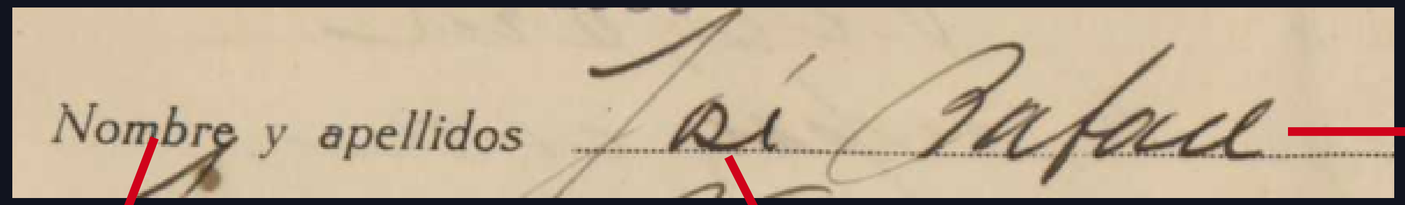

Most likely hypotheses
Word
Condidence
REGISTADO
55.50%
REGUNDO
42.50%
SEGUINDO
34.40%
SEGISTADO
29.40%
REGISTO
24.10%
SEGUNDO
17.60%
REGUINDO
10.10%

Transcription + Alternatives



November 2023

Metadata Extraction



Most probable Hypotheses

Nombre!impr

Add

Word	Confidence	GT	X
Nombre!impr	100.00%	0	X

Most probable Hypotheses

José!nombre

Add

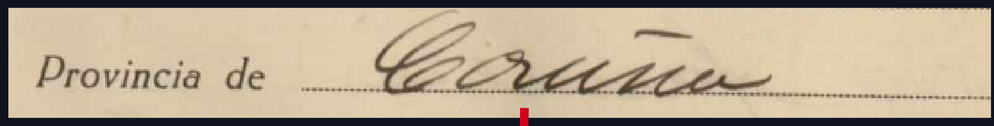
Word	Confidence	GT	X
José!nombre	97.90%	0	X

Most probable Hypotheses

Rafael!nombre

Add

Word	Confidence	GT	X
Rafael!nombre	53.60%	0	X

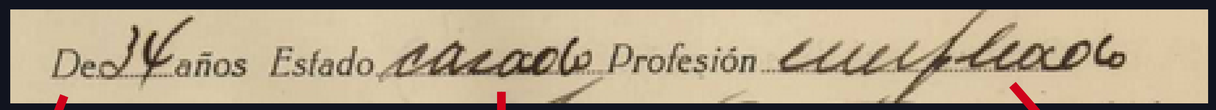


Most probable Hypotheses

Coruña!provincia

Add

Word	Confidence	GT	X
Coruña!provincia	100.00%	0	X



Most probable Hypotheses

34!edad

Add

Word	Confidence	GT	X
34!edad	100.00%	0	X

Most probable Hypotheses

casado!estadocivil

Add

Word	Confidence	GT	X
casado!estadocivil	100.00%	0	X

Most probable Hypotheses

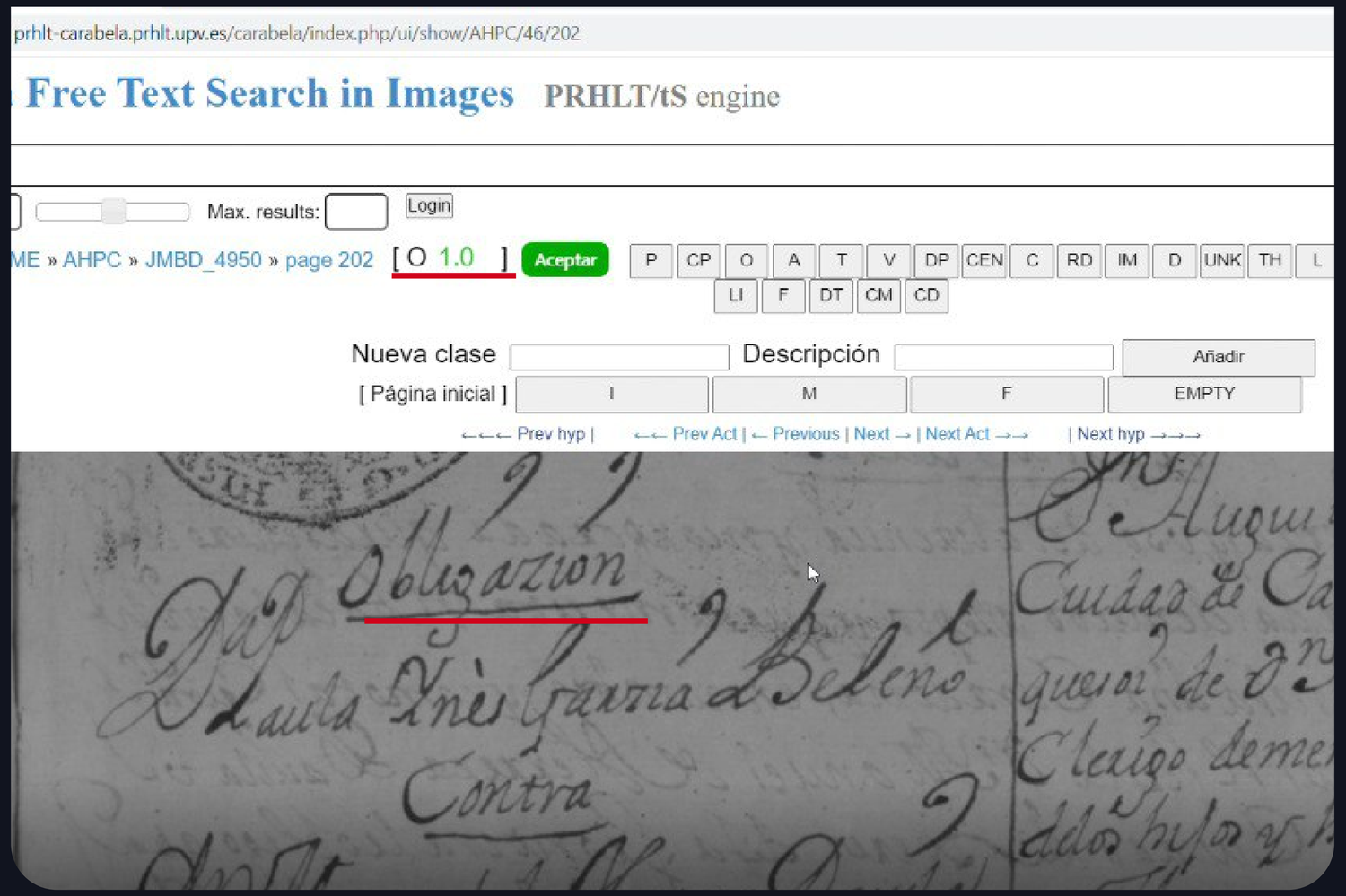
empleado!oficio

Add

Word	Confidence	GT	X
empleado!oficio	100.00%	0	X

Anonymization

Classification & Segmentation



Big Data Analysis

Civil State

		0%
1096.8	soltero	36%
1015.7	casado	34%
		70%
437.0	soltera	14%
209.0	casada	7%
		91%
120.5	viuda	4%
70.3	viudo	2%
21.7	célibe célibero	1%
		98%
62.2	[OTHER and ERRORS]	2%
		100%

Genre

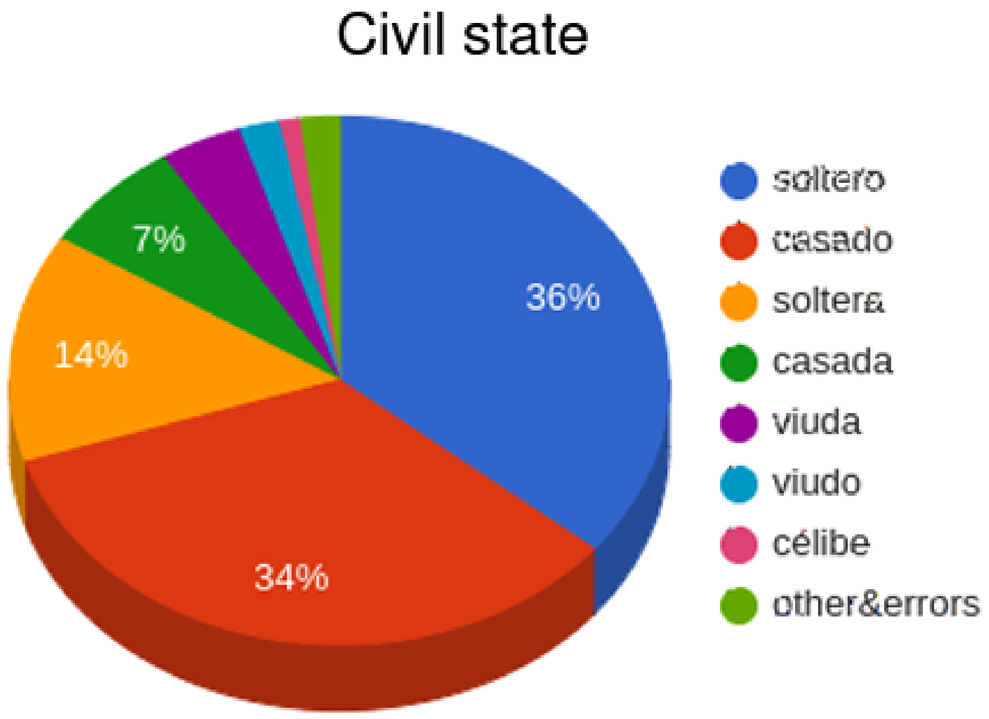
2182.8	Men:	72%
766.5	Women:	25%
73.9	Unknown:	3%

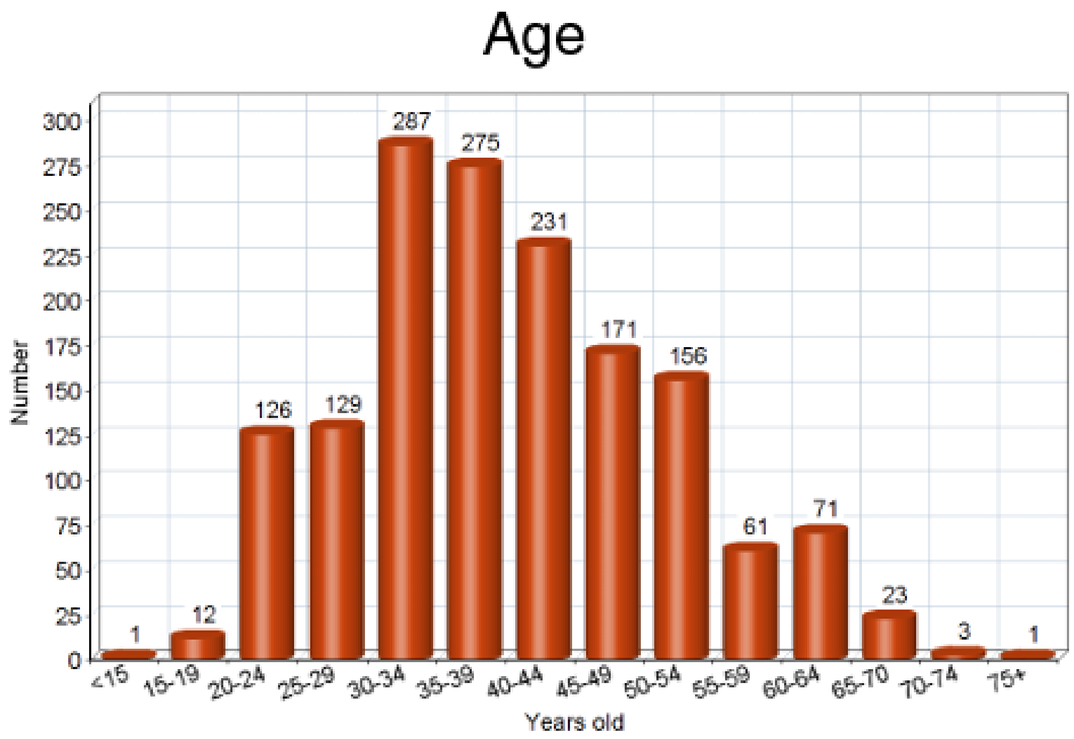
Age (years old)

		0%
1.4	<15	
11.9	15-19	
		10%
126.3	20-24	
129.0	25-29	
286.6	30-34	
		50%
275.0	35-39	
230.5	40-44	
170.7	45-49	
156.2	50-54	
		90%
61.2	55-59	
71.4	60-64	
		99%
23.1	65-70	
3.2	70-74	
1.1	75+	
		100%

November 2023

Big Data Analysis





November 2023

Demonstration

Thank you!

www.transkriptorium.com

info@transkriptorium.com