A Language Modelling Approach to Quality Assessment of OCR'ed Historical Text

Callum W Booth, Robert Shoemaker, Robert Gaizauskas
University of Sheffield, Sheffield, United Kingdom
{cwbooth1, r.shoemaker, r.gaizauskas}@sheffield.ac.uk

The
University
Of
Sheffield.

Part of project: "Information Extraction and Entity Linkage in Historical Crime Records"

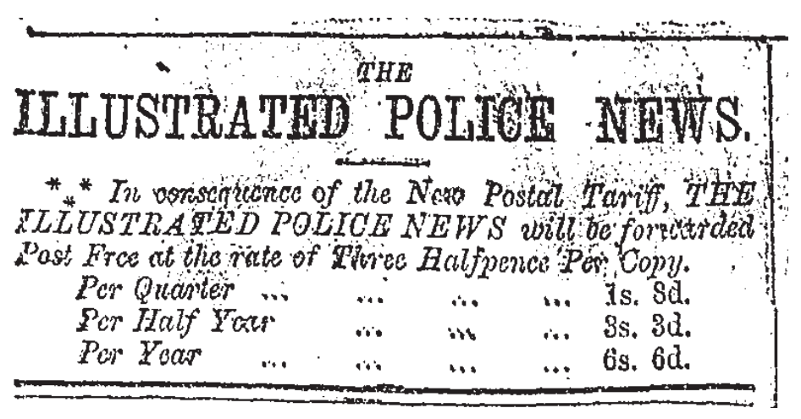
INTRODUCTION AND OUTLINE

Goal: Reduce a dataset of newspaper transcriptions down to the best quality OCR.

Project Goal: Capture details of criminal lives from nineteenth century newspaper reports, and link them to the Digital Panopticon.

Limitations:

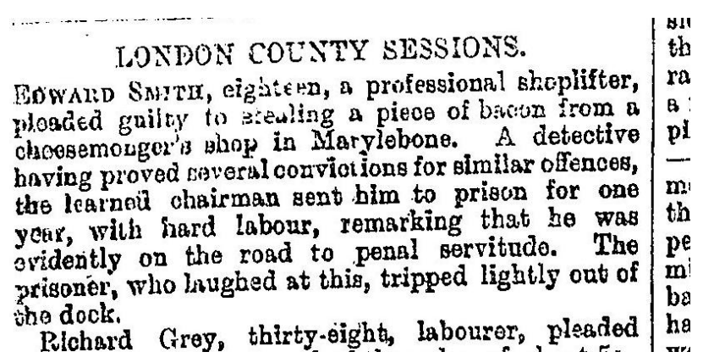
- Gold-standard transcriptions are not available.
- OCR quality measurement is more commonly viewed as an intrinsic evaluation method for OCR systems themselves. [1, 2]
- OCR quality within the dataset is highly variable, due to various scan artefacts.



THE
ILLUSTRATED POLICE NEWS.
*** In consequence of the New Postal Tariff, THE ILLUSTRATED POLICE NEWS will be forwarded Post Free at the rate of Three Halfpence Per Copy.

Per Quarter	1s. 8d.
Per Half Year	3s. 3d.
Per Year	6s. 6d.

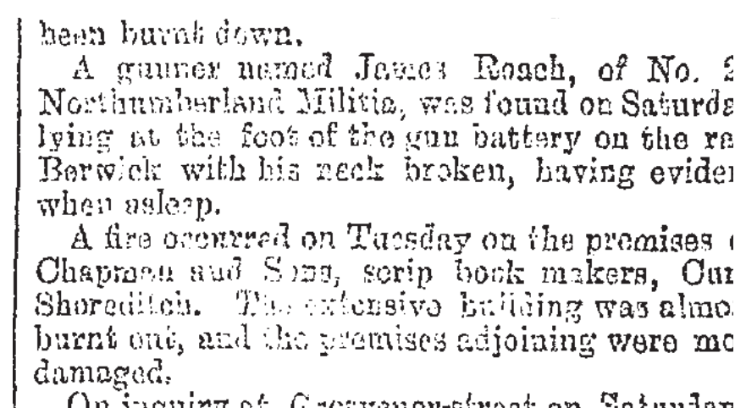
Doc BA3200797019 [3] - Example of debris or ink run/bleed.



LONDON COUNTY SESSIONS.

Edward Smith, eighteen, a professional shoplifter, pleaded guilty to stealing a piece of bacon from a cheesemonger's shop in Marylebone. A detective having proved several convictions for similar offences, the learned chairman sent him to prison for one year, with hard labour, remarking that he was evidently on the road to penal servitude. The prisoner, who laughed at this, tripped lightly out of the dock.
Richard Grey, thirty-eight, labourer, pleaded

Doc BA3200808407 [4] - Example of rotation and overlap from adjacent columns.



been burnt down.
A gunner named James Roach, of No. 2 Northumberland Militia, was found on Saturda lying at the foot of the gun battery on the ra Berwick with his neck broken, having evide when asleep.
A fire occurred on Tuesday on the premises Chapman and Sons, scrip book makers, Cu Shoreditch. The extensive building was almo burnt out, and the premises adjoining were mo damaged.

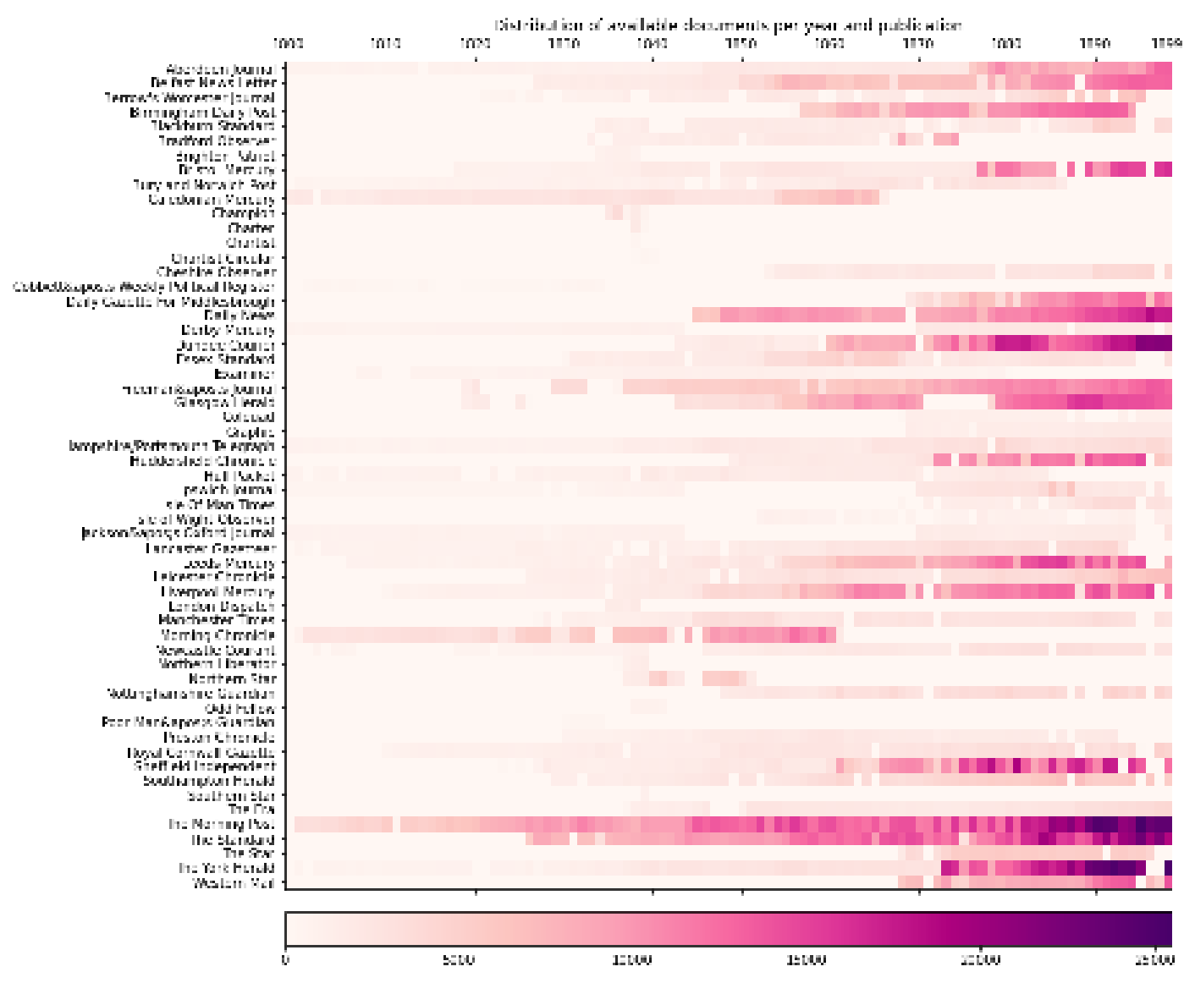
Doc BA3200803387 [5] - Example of ink fading and general damage.

DATASETS AND INITIAL CORPUS ANALYSIS

We use two main datasets: the British Library Newspapers [BLN] dataset, and the Proceedings of the Old Bailey Online [OBP] dataset.

BLN

- Transcribed by Gale from British Library microfilm.
- Parts 1 and 2 comprises over 14 million article transcriptions from nineteenth century British newspapers.



OBP

- A digitised collection of trial reports from London's central criminal court.
- Collections of reports were regularly published whenever the court met (between 8 and 12 times a year). [6]
- Scans were transcribed through either double-rekeying, or comparison of a single-keying and a machine transcribed copy. [7]
- Transcription accuracy of "well over 99%". [7]

MODEL ARCHITECTURE

- We fashion a language model trained on texts from the Proceedings of the Old Bailey Online.
- We exploit the genre adjacency between Old Bailey trial reports and newspaper crime reports.
- We create a weighted ensemble model of bigrams, unigrams, and a uniform zeroth order model for smoothing.
- Probabilities are estimated using MLE.

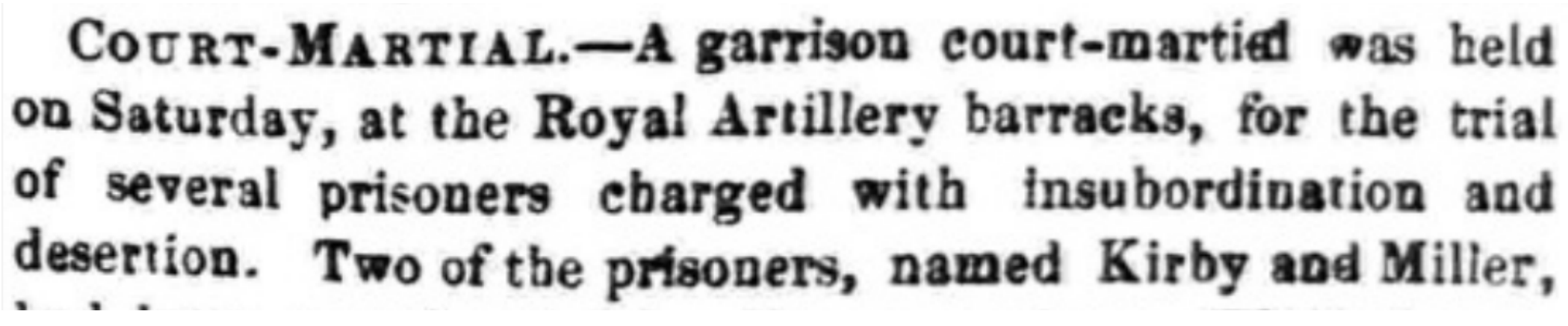
$$\begin{aligned} P(w_k|w_{k-1}) = &\lambda_1 P_{\text{bigram}}(w_k|w_{k-1}) + \\ &\lambda_2 P_{\text{unigram}}(w_k) + \\ &\lambda_3 |\mathbf{V}|^{-1} \end{aligned}$$

- Seperate models are trained per decade of OBP texts to account for historical changes in legal parlance. [8]

MODEL EVALUATION

- To score and rank OCR quality, we compute and sort by average log likelihood for each document in a London-specific corpus of 17 publications against its matching decade model.
- We locate the first crime report within various percentiles of the ranked dataset and manually verify the transcription quality.

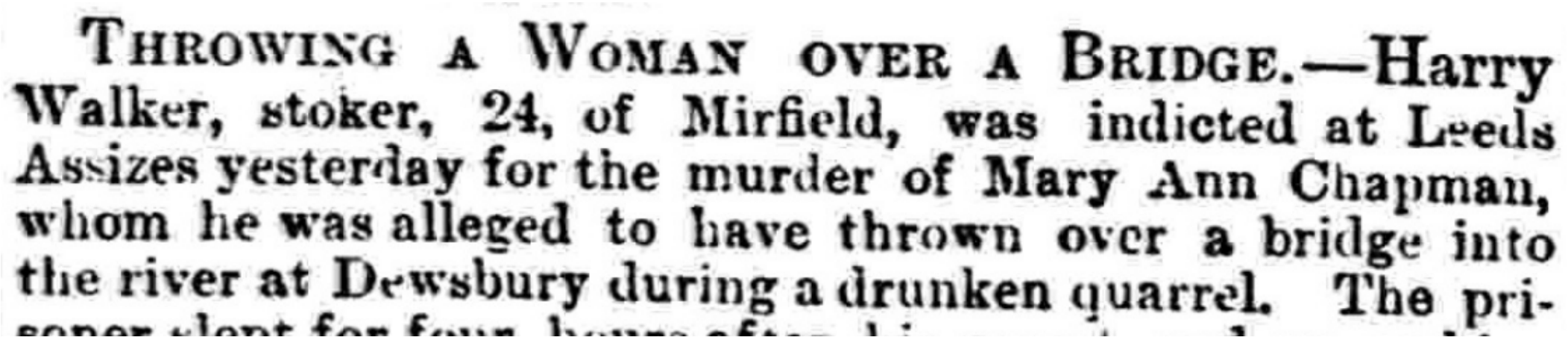
1st percentile - [9]



Court-Martial.—A garrison court-martial was held on Saturday, at the Royal Artillery barracks, for the trial of several prisoners charged with insubordination and desertion. Two of the prisoners, named Kirby and Miller,

A garrison court-martial was beld on Saturday, at the Royal Artillery barracks, for the trial of several prisoners charged with insubordination and desertion.
100% correct entities, 96% correct tokens

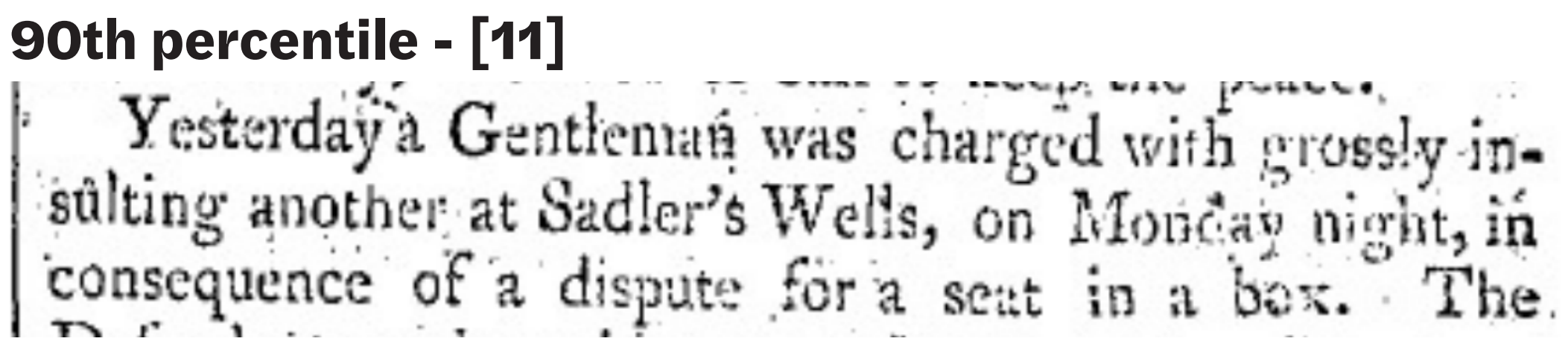
10th percentile - [10]



Throwing a Woman over a Bridge.—Harry Walker, stoker, 24, of Mirfield, was indicted at Leeds Assizes yesterday for the murder of Mary Ann Chapman, whom he was alleged to have thrown over a bridge into the river at Dewsbury during a drunken quarrel. The pri-

Harry Walker, stoker, 24, of Mirfield, was indicted at Leeds Assizes yesterday for the murder of Mary Ann Chapman, whom he was alleged to have thrown over a bridge into the river at Dewsbury during a drunken nuarreL
100% correct entities, 97% correct tokens

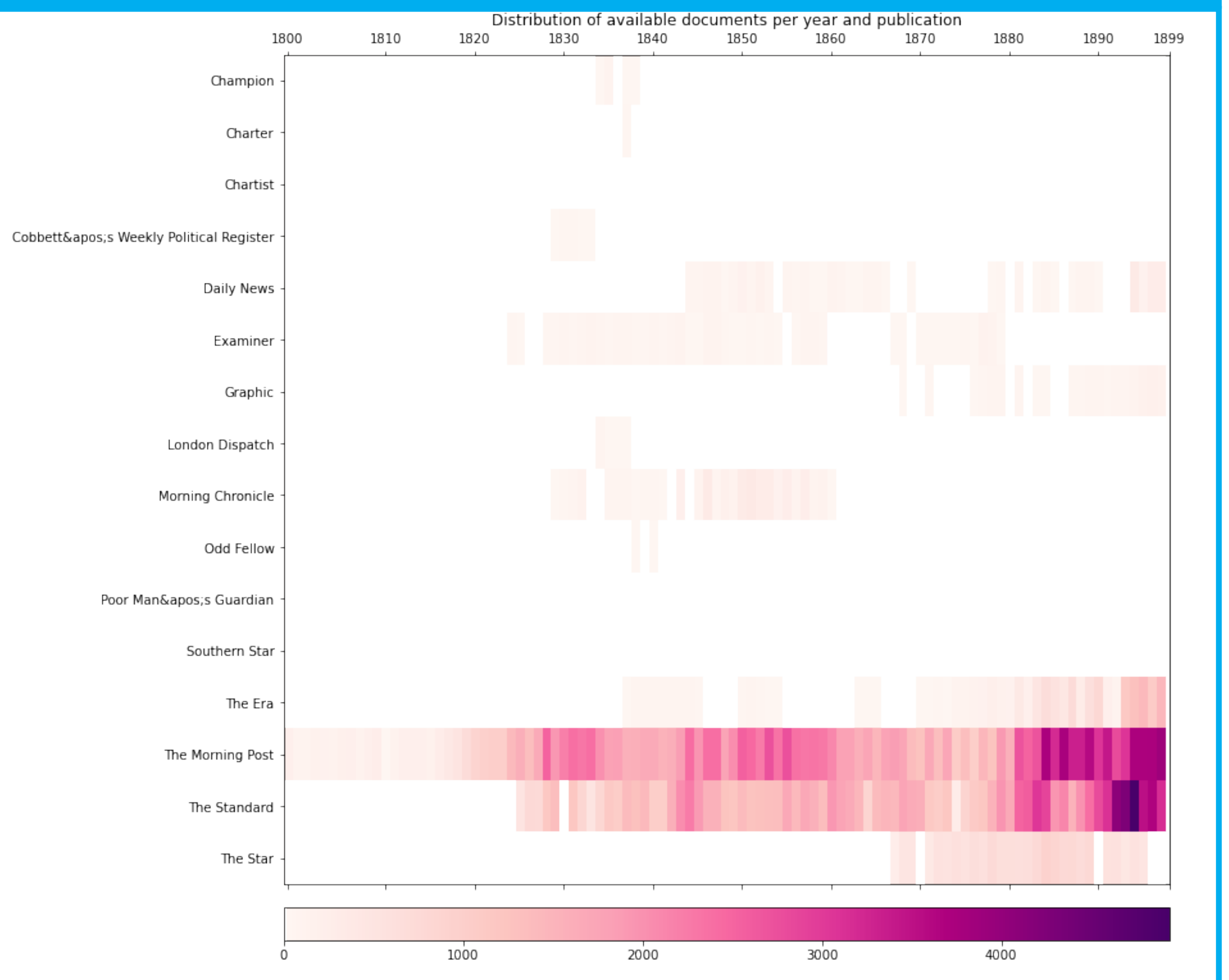
90th percentile - [11]



Yesterday a Gentleman was charged with grossly insulting another at Sadler's Wells, on Monday night, in consequence of a dispute for a seat in a box. The

' Yesterdaj't Gentleniani was charged vi!vh urossly in. fdulting another at Sadler't Wells, on ui i n;t, in consequence of Pa dispute .fdr a seat in o b);
0% correct entities, 50% correct tokens

FINAL DATASET ANALYSIS AND CONCLUSIONS



- The initial corpus restriction to London-specific publications represents a ~76% reduction from 14 million documents to 3.3 million.
- Using this language model methodology, we select the top 10% of documents ranked by OCR quality to form the final working corpus.
- The final corpus comprises 338k documents, a ~97% reduction from the initial corpus.
- These documents represent the best quality transcriptions of London-specific newspaper articles.
- The final corpus maintains skew towards the late nineteenth-century, and towards more document dense publications, such as *The Morning Post* and *The Standard*, which constitute 87% of the working corpus, an increase from their 17% initial corpus share.
- We conclude that we can use language modelling techniques in conjunction with adjacent genre datasets to measure and rank quality of OCR'ed historical documents.

REFERENCES

[1] Rice, S. V. (1996). Measuring the Accuracy of Pagereading Systems. Ph.D. thesis, University of Nevada, Las Vegas, NV.

[2] Nartker, T., Rice, S., and Lumos, S. (2005). Software Tools and Test Data for Research and Testing of Pagereading OCR Systems. In Document Recognition and Retrieval XII, pages 37–47. International Society for Optics and Photonics, SPIE.

[3] BLN. (1882). Advertisements & Notices. Illustrated Police News, (954), 27 May. 1882. link-gale-com.sheffield.idm.oclc.org/apps/doc/BA3200797019/BNCN?u=su_uk&sid=bookmark-BNCN&xid=91ec0f11.

[4] BLN. (1889). LONDON COUNTY SESSIONS. Illustrated Police News, (1343), 9 Nov. 1889. link-gale-com.sheffield.idm.oclc.org/apps/doc/BA3200808407/BNCN?u=su_uk&sid=bookmark-BNCN&xid=3a9265ad.

[5] BLN. (1886). Everybody's Column. Illustrated Police News, (1175), 21 Aug. 1886. link-gale-com.sheffield.idm.oclc.org/apps/doc/BA3200803387/BNCN?u=su_uk&sid=bookmark-BNCN&xid=0a25d08e.

[6] Emsley, C., Hitchcock, T., and Shoemaker, R. (2018). The Proceedings - Publishing History of the Proceedings, Old Bailey Proceedings Online. https://www.oldbaileyonline.org/static/Publishinghistory.jsp. version 8.0.

[7] Emsley, C., Hitchcock, T., and Shoemaker, R. (2018). Old Bailey Online - About This Project, Old Bailey Proceedings Online. https://www.oldbaileyonline.org/static/Project.jsp. version 8.0.

[8] Emsley, C., Hitchcock, T., and Shoemaker, R. (2018). Crime and justice - Crimes Tried at the Old Bailey, Old Bailey Proceedings Online. https://www.oldbaileyonline.org/static/Crimes.jsp. version 8.0.

[9] BLN. (1858). Multiple News Items. Morning Post, (26291):3, 5 Apr. 1858. link-gale-com.sheffield.idm.oclc.org/apps/doc/R3213129190/BNCN?u=su_uk&sid=bookmark-BNCN&xid=52eb3625.

[10] BLN. (1892). Multiple Classified ads. Morning Post, (37427):6, 27 May. 1892. link-gale-com.sheffield.idm.oclc.org/apps/doc/R3214411435/BNCN?u=su_uk&sid=bookmark-BNCN&xid=3007357b.

[11] BLN. (1805). POLICE. Morning Chronicle [1801], (11349), 2 Oct. 1805. link-gale-com.sheffield.idm.oclc.org/apps/doc/BB3207096330/BNCN?u=su_uk&sid=bookmark-BNCN&xid=f3a87044.