

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Estimating and Rating the Quality of Optically Character **Recognised Text**

Citation for published version:

Alex, B & Burns, J 2014, Estimating and Rating the Quality of Optically Character Recognised Text. in Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage. ACM Association for Computing Machinery, Madrid, Spain, pp. 97. https://doi.org/10.1145/2595188.2595214

Digital Object Identifier (DOI):

10.1145/2595188.2595214

Link: Link to publication record in Edinburgh Research Explorer

Published In: Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Estimating and Rating the Quality of Optical Character Recognised Text



Beatrice Alex <u>balex@inf.ed.ac.uk</u>



OVERVIEW

- Background: Trading Consequences
- OCR accuracy estimation
 - Motivation

- Related work
- OCR errors in text mining (eye-balling data versus quantitative evaluation)
- Computing text quality
- Manual vs. automatic rating
- Summary and conclusion

TRADING CONSEQUENCES

- JISC/SSHRC Digging into Data Challenge II (2 year project, 2012-2013)
- Text mining, data extraction and information visualisation to explore big historical datasets.
- Focus on how commodities were traded across the globe in the 19th century.
- Help historians to discover novel patterns and explore new research questions.

PROJECT TEAM





Ewan Klein, Bea Alex, Claire Grover, Richard Tobin: *text mining*

Colin Coates, Andrew Watson: historical analysis



Jim Clifford: historical analysis



James Reid, Nicola Osborne: *data management, social media*

Aaron Quigley, Uta Hinrichs: *information visualisation*

TRADITIONAL HISTORICAL RESEARCH

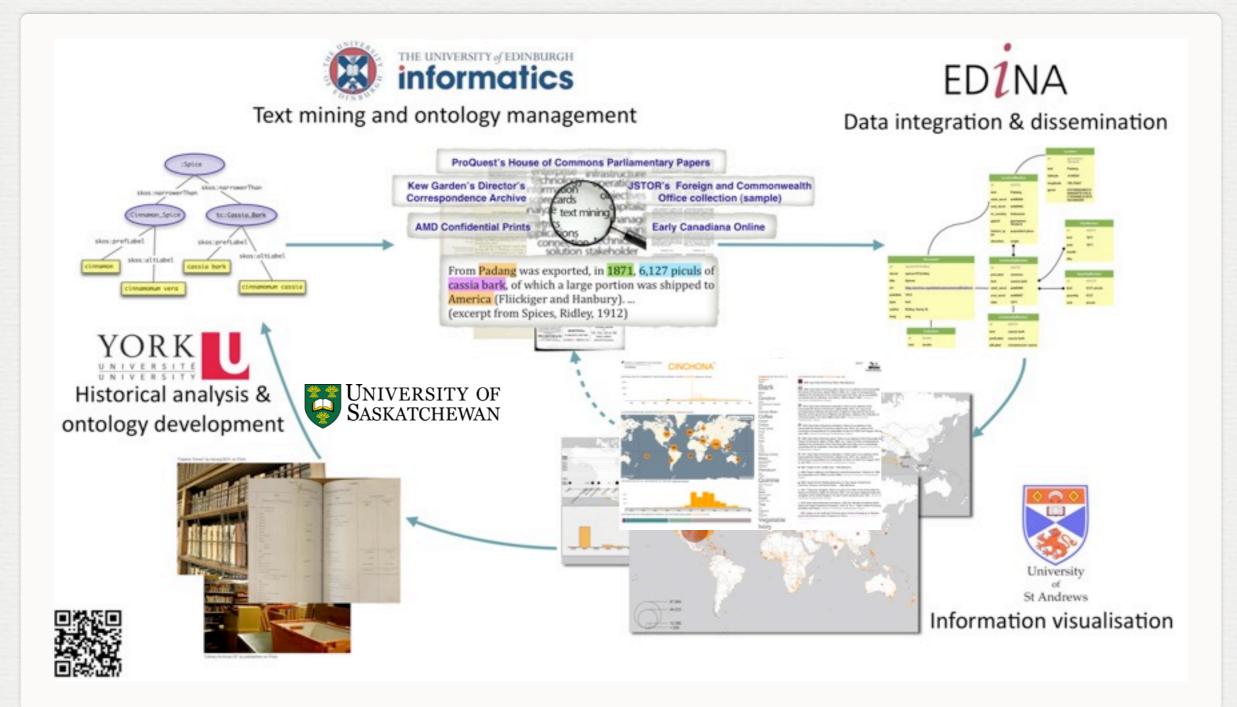


Map showing the areas where mahogany is grown

Gillow and the Use of Mahogany in the Eighteenth Century, Adam Bowett, Regional Furniture, v.XII, 1998.



PROJECT OVERVIEW



Scotland's National Collections and the Digital Humanities, Edinburgh, 14/02/2014

DOCUMENT COLLECTIONS

Collection	# of Documents	# of Images
House of Commons Parliamentary Papers (ProQuest)	118,526	6,448,739
Early Canadiana Online	83,016	3,938,758
Directors' Letters of Correspondence (Kew)	14,340	n/a
Confidential Prints (Adam Matthews)	1,315	140,010
Foreign and Commonwealth Office Collection	1,000	41,611
Asia and the West (Gale)	4,725	948,773 (OCRed: 450,841)

Scotland's National Collections and the Digital Humanities, Edinburgh, 14/02/2014

DOCUMENT COLLECTIONS

Collect	tion	# of Documents	# of Images					
House for Parlia Early (Over 10 r	nillion docume 7 billion word t	ent pages,					
Asia and the V	Asia and the West (Gale) 4,725 948,773 (OCRed: 450,841)							

Scotland's National Collections and the Digital Humanities, Edinburgh, 14/02/2014

OCR-ED TEXT

<?xml version="1.0" encoding="UTF-8"?>

<article id="10.2307/60227644">

<page> <![CDATA[T, G1IXET, Printer.]]> </page>

cpage> <![CDATA[' INTRODUCTION, AS the following Vindication may fall into the hands of perfons who have</pre> never read the Hiftory of the Politicks of Great Britain and France, it will not be improper, before I enter on my Defence, to ftate the principal facts, which were fuccef- fively proved by authentic documents, in the fixteen chapters, of which that wrork is compofed. - r 1. In the celebrated conference at PiUnitz; in Auguft, i;gi, the Britifh Government took not the rnoft diftant part: and if.-any treaty was concluded there, which is itfelf a matter of great doubt, the Britifh Go- vernment not only never acceded to it, but was, never apprifed even of its contents.- Further, when the Britiih Government was requefted in 1701 to join a coalition againft France, it gave a pofitive and unequivocal refufal. B 2 2. Toward]]> </page> colony of St. Domingo was pre- served to France by the timely affiftance fent by Lord Effingham, then Governor of Jamaica : and the Britifh. Cabinet fignified through its AmbafTador at Paris to the French Government, that it fully approved of Lord Effingham's conduct.. At the fame time, true to the ftri&eir. principles of ho- nour and neutrality, it refufed the advan- tageous offer made by the French colonifts, who were highly diflatisfied with the National AfTembly, to furrender the French part of St. Domingo to the Crown of Bri- tain. And thefe a6ls of generofity were re* paid by France with the utmoft ingrati- tude. 3. When Louis XVI. formally accepted the new conflitution, in September, 17Q1, and fent circular letters to the different Courts of Europe fignifying his affent, the Court of Great Britain was one of the firft which returned an anfwer ; and the anfwer was couched in very refpectful terms, where- as fome other courts either did not anfwer at HfWta]]> </page> ...

</article>

OCR-ED TEXT

<?xml version="1.0" encoding="UTF-8"?>
<article id="10.2307/60227644">
</article id="10.2307/60227644"</article id="10.2307/60227644">
</article id="10.2307/60227644">
<

Proclamations, Pro * v mie RL' E.LI S B.AIG07. iVICTORIFIA. h>I 1/ t(aŤ'' of' GO!>. tif ih Firi. ea fil~~Ť/ r<' lluil'tIŤ, (i'. i', QUEE'. Tc ,iiin iŤiV i ' ui tillhŤ'nt, 111te 1 eihŤ' Colin. ('it;ZI-s. uni 14Lt1ussuls ce t rib ev iii tJ1u stat. it have Iei t's.iiititŤud ztntt liild, a tutt < A 11i10C.

generofity were re* paid by France with the utmoft ingrati- tude. 3. When Louis XVI. formally accepted the new conflitution, in September, 17Q1, and fent circular letters to the different Courts of Europe fignifying his affent, the Court of Great Britain was one of the firft which returned an anfwer; and the anfwer was couched in very refpectful terms, where- as fome other courts either did not anfwer at HfWta]]> </page>

OCR-ED TEXT

<?xml version="1.0" encoding="UTF-8"?>
<article id="10.2307/60227644">
<page> <![CDATA[THE HISTORY OP THE POLITICKS OF GREAT BRITAIN AND FRANCE, VINDI GATED FROM A LATE ATTACK OF
MR. WILLIAM BELSHAM. BY HERBERT MARSH, B. D. F. R. S. and tellow or st. John's college, Cambridge. Ecntmn:
PRINTED FOR JOHN STOCKDALE, PICCADILLY. 1801. t35)lvjf~ Udf4~ P.]]> </page>

qBiu si }S3A:req s,uauuaqsu aq} }Bq} uirepo.ifT 'papua}X3 sSuiav }qSuq Jiaq} qiiM jib ui snnS bbs aqx 'a"3(s aq} tnojj ssfitns q}TM Sni5[ooi si jb}s }S.ii; aqx 'papnaoSB q}Bq naABSjj qS;H °1 ssbui s.uauuaqsu aqx

Extract from document 10.2307/60238580 in FCOC.

couched in very refpeclful terms, where- as fome other courts either did not anfwer at HfWta]]> </page>
...|
</article>

aqx JX u aqx

)F

Thou dost enre. Thon art there.

not receive as blessed dower Even in the darkest hour, We will own Thy mighty power, And can still. Thou didst once its wrath assuage, Is Thy power from age to age ; Stronger than the tempest's rage

All Thy will.

And the waves leap to the sky, Wherefore should we fret and sigh ? When the storm is raging high,

When upon the stormy deep,

Again and again her look upward renewing,

Her baby's soft check, as the rosebud as fair; Again and sessid buot dity niggs has night

Fishermen their vigils keep, Wherefore should their lovers weep?

This song, sweet and solemn, she breathes on the air :

Proclaim that the fishermen's harvest is nigh. . The sea gulls in air with their bright wings extended, 'The first star is looking with smiles from the sky, The fishermen's mass to High Heaven bath ascended,

As shone from each face in the bay of Tralee. worrom-of raddgird ni dist dous rot b'yarq buA

Sing some soon to form for the brow of the sun? Watching the brooklets their pearly drops squander, Where quiet and beauty are blended in one, By fair Tralee bay, oh! did you e'er wander,

And heard their low throbs of worshipping feeling. Their faces turned upward beneath the bright sky, Oh 1 I have been there, and seen them all kneeling, ? selarT tsiup to eldmud memorian adT As mass hath been said by those some of the ocean, Like chorus of angels from depths of the sea, And have you e'er heard the swell of devotion,

I've thought of that fishermen's mass of the sea, And off, 'mid the worlds rough battles and sources, And seen pure devotion guan out from each eye;

pap

<?>

<ar

< p

MR.

PRI

Befriend !--Surve, the fishermen's mass it was said." Cries " Mother of Jesus, and saints the most holy, And, wiping the brine from his storm-beaten head, Brave Dermot looks up with a spirit all lowly.

Fast bugging her hehe to her bosom high swelling, Tight clasping her rosary, Mary's low bowed. There allines a faint light like a star through a cloud; Upon the dark storm, from the fisherman's dwelling,

Then, hissing her tenrs to her Father in heaven, Then, hissing her child, cries, "Ah, bless'd be that

Sure, mass it was said-he'll he here in a while. I novig ai saimorq add won I stains add ad b'ssaid buA i alime

OCR-ED TEXT

The Fishermen's Mass.

Extrac 10.230

COL

.. <1

WHY OCR ACCURACY ESTIMATION?

- A reasonable amount of already digitised books (some with very bad text quality). Can we mine some of them now.
- To what extent do OCR errors affect text mining? What is their effect when dealing with big data?
- What text is of sufficient high quality to be understood? How bad is too bad? What happens to the rest?
- Can we measure text quality? How does it compare to human quality ranking of text?

RELATED WORK

- Some OCR output contains character-based accuracy rates which can be very deceptive.
- Popat, 2009:
 - Extensive study on quality ranking of short OCRed text snippets in different languages. Examined rank order of text snippets of inter-, intra- and machine ratings.
 - Compared spatial and sequential character n-gram-based approaches to a dictionary-based approach (web corpus, capped at 50K most frequent words per language).
 - Compared random to balanced (stratified) sampling.
 - Metric: average rank correlation.

OCR ERRORS AND BIG DATA

- Are OCR errors negligible when mining big data to detect trends?
- Our data suffers from all the common OCR error types (at best just a few character insertions, substitutions and deletions), at worst much worse (page upside down).
- Character confusion examples:
 - e -> c, a -> o, h -> b, l -> t, m -> n, f -> s





Google books Ngram Viewer

Graph these cas		omma-separated p			with smoothing	of 3 🛟.		Q⁺ Shan Ƴ Twe	50
.008 Search lots of b									
.007									
0.000002% 0.0005	mohogan	y 📕 mabogan	У						
0.0004 0.0000016%									
0.0003 0.0000012%									
0.000008% 0.001								3	
0.0000004%				Λ		7 0	6	man	
0.00%	1680 170	0 1720 1740	1760 1780	1800 1820	1840 1860	1880 190	0 1920 1	940 1960	1980 2000

Google books Ngram Viewer

between 1730 and	2000 from the corpus	English	with smoothing o	f 3 🛟.	Tweet 0
B Search lots of books					
)7					
06	hogany 📕 maboga	ny 📕 mahogany			
0.0004%				Δ	
04 0.00032%				1 h	\sim
0.00024%		6 6			
02			m	m	M
0.00016%		M mm			
0.0008%		In			
0.00%					

Google books Ngram Viewer

veen 1700 and 2000	from the corpus	English	with smo	oothing of	3 🗘 .		Tweet	0
arch lots of books								
time	lime							
0.15%								
0.129/	Jun	m						
0.12%								
0.09%	4					8		
0.06%	<i>V</i> .							
0.03%								

OCR ERRORS AND TEXT MINING

- Need a more quantitative analysis.
- Built a commodity and location recognition tool.
- Evaluated it against manually annotated gold standard.

	\mathbf{TP}	\mathbf{FP}	\mathbf{FN}	Р	R	F
TM: com	797	342	310	0.70	0.72	0.71
TM: loc	1,599	489	$1,\!549$	0.77	0.51	0.61
IAA: com	283	112	109	0.72	0.72	0.72
IAA: loc	582	65	189	0.90	0.76	0.82

OCR ERRORS AND TEXT MINING

- 32.6% of false negative commodity mentions (101 of 310) contain OCR errors (= 9.1% of all commodity mentions in the gold standard)
 - sainon, rubher, tmber
- 30.2% of false negative location mentions (467 of 1,549) contain OCR errors (= 14.8% of all location mentions in the gold standard)
 - Montreai, Montroal, Mont- treal and 10NTREAL.

OCR ERRORS AND TEXT MINING

9, Montreai 2, Montroal 2, Montrent 2, Montrea 1, MO.'N'YREUL 1, Mont- treal 1, MONTRLAL 1, Montreali 1, MONTREAL 1, Mont real 1, MONTRBf'tL 1, MONTIIEAL 1, MIontret] 1, Mbontreal 1, Maontreal 1, 3MON2RRA 1,10TRBAL 1,10NTREAL

PREDICTING TEXT QUALITY

- Can we compute a simple quality score for a large data collection (i.e. over 7 billion words)?
- How easily can humans perform document-level quality rating?

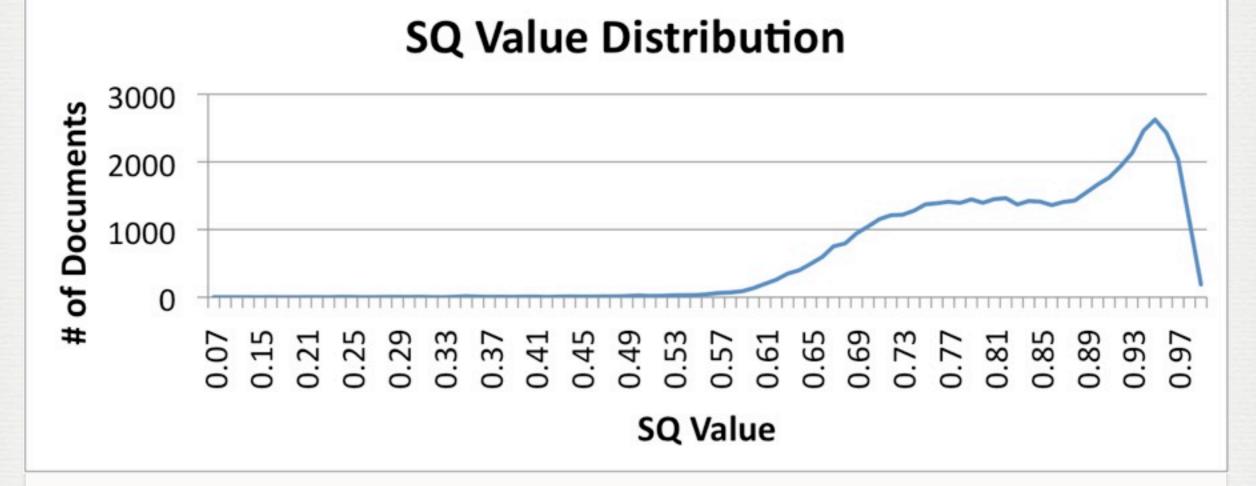
COMPUTING TEXT QUALITY

- Simple document-level quality score to get a rough estimate of how good a document is.
- Word tokens found in an English dictionary (aspell "en") and Roman/Arabic numbers over all word tokens in the text.
 - Scores range between 0 and 1.
- Caveat: it does not consider historic variants.

$$SQ = \frac{W_{good}}{W_{all} + 1}$$

COMPUTING TEXT QUALITY

 Score distribution over the English Early Canadiana Online data (55,313 documents).



DATA PREPARATION

- Early Canadiana Online (books, magazines and government publications relevant to Canadian history ranging from 1600 to the 1940s)
- 83,016 documents (almost 4 million images containing text mostly in English and French but also in 10 First Nation languages, European languages and Latin).
- Language identification (or meta data information) to retain only English content (55,313 documents).

DATA PREPARATION

- Ran the automatic scoring over all English ECO documents.
- Applied stratified sampling to collect 100 documents by randomly selecting:
 - 20 documents where $0 \ge SQ < 0:2$,
 - 20 documents where $0.2 \ge SQ < 0.4$,
 - 20 documents where $0.4 \ge SQ < 0.6$,
 - 20 documents where $0.6 \ge SQ < 0.8$,
 - 20 documents where $0.8 \ge SQ < 1$.
- Shuffled documents and removed the quality score.

MANUAL RATING

Two raters looked at each document and rated it on a 5-point scale.

5...OCR quality is high. There are few errors. The text is easily readable and understandable.

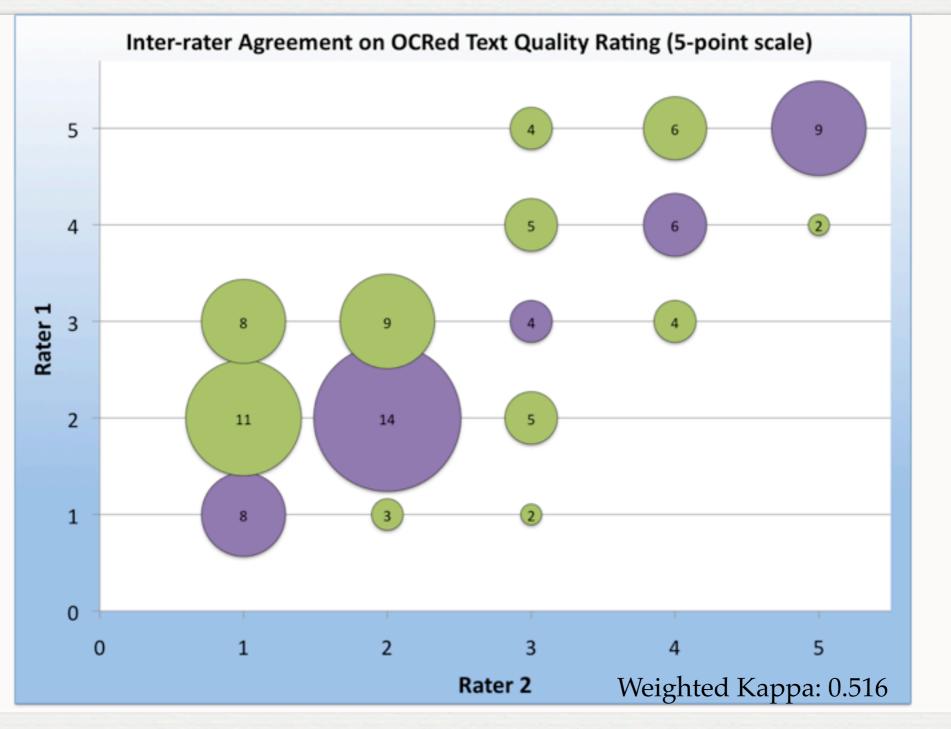
4 ... **OCR quality is good.** There are some errors but they are limited in number and the text is still mostly readable and understandable.

3 ... OCR quality is mediocre. There are numerous OCR errors and only part of the text is readable and understandable.

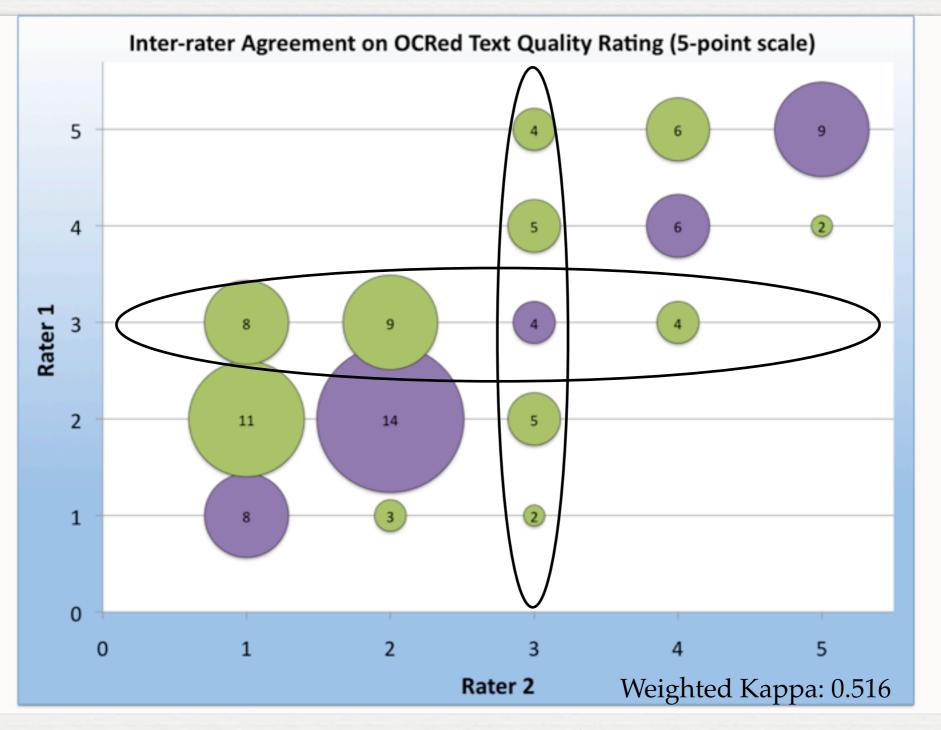
2 ... OCR quality is low. There is a large number of OCR errors which seriously affect the readability and understandability of the majority of the text.

1 ... OCR quality is extremely low. The text is so full of errors that it is not readable and understandable.

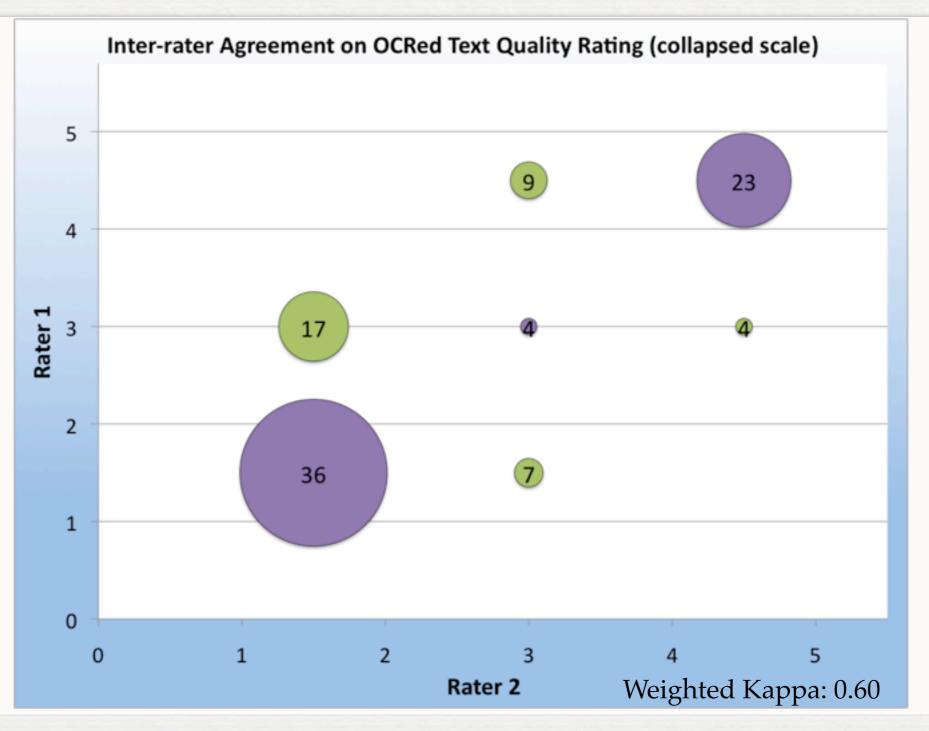
INTER-RATER AGREEMENT



INTER-RATER AGREEMENT

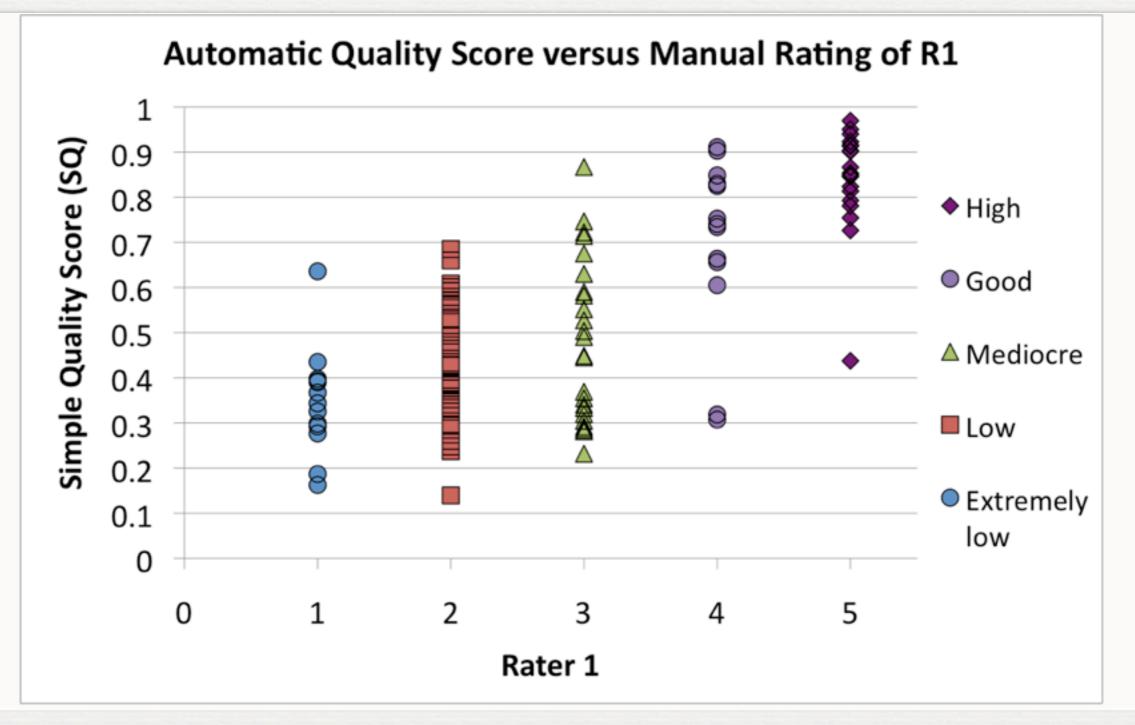


INTER-RATER AGREEMENT

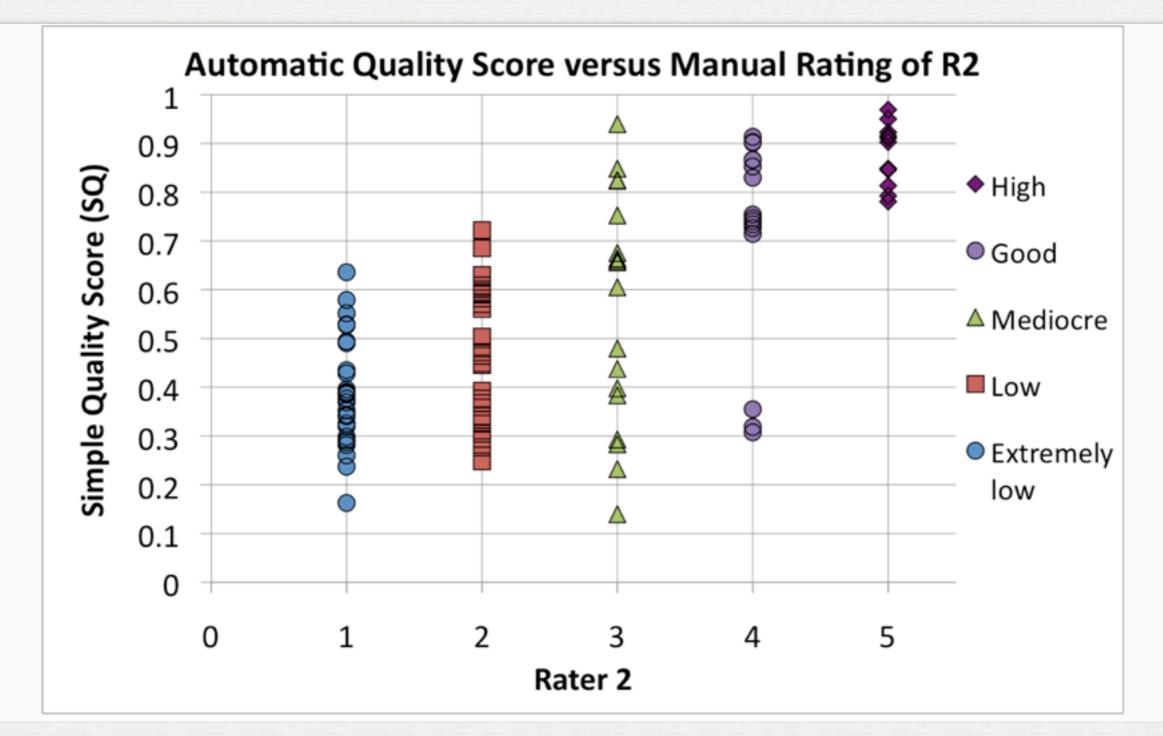


DATeCH 2014, May 20th 2014

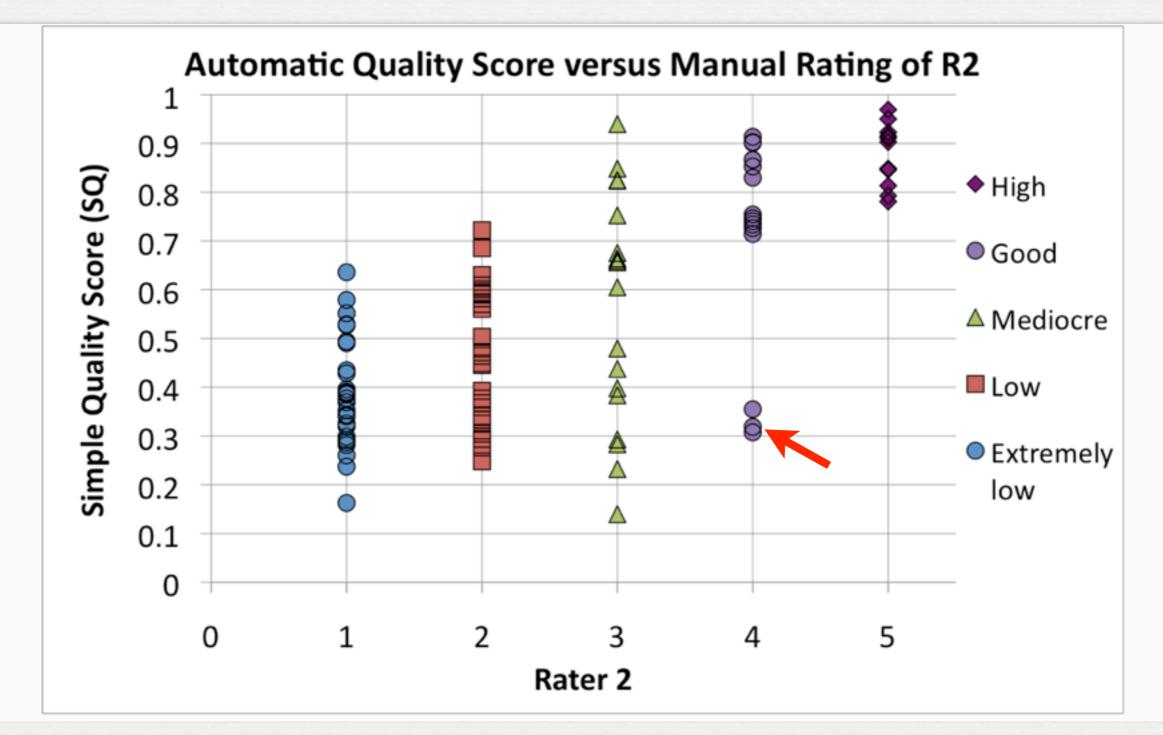
AUTOMATIC VS HUMAN



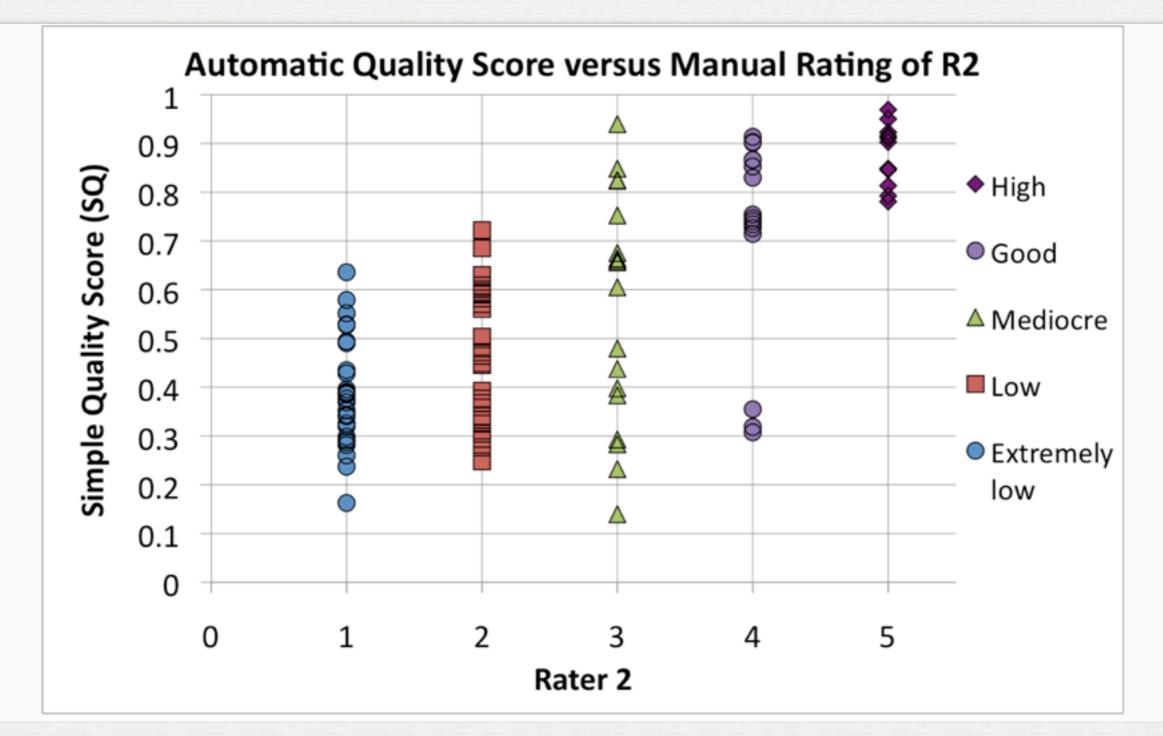
AUTOMATIC VS HUMAN



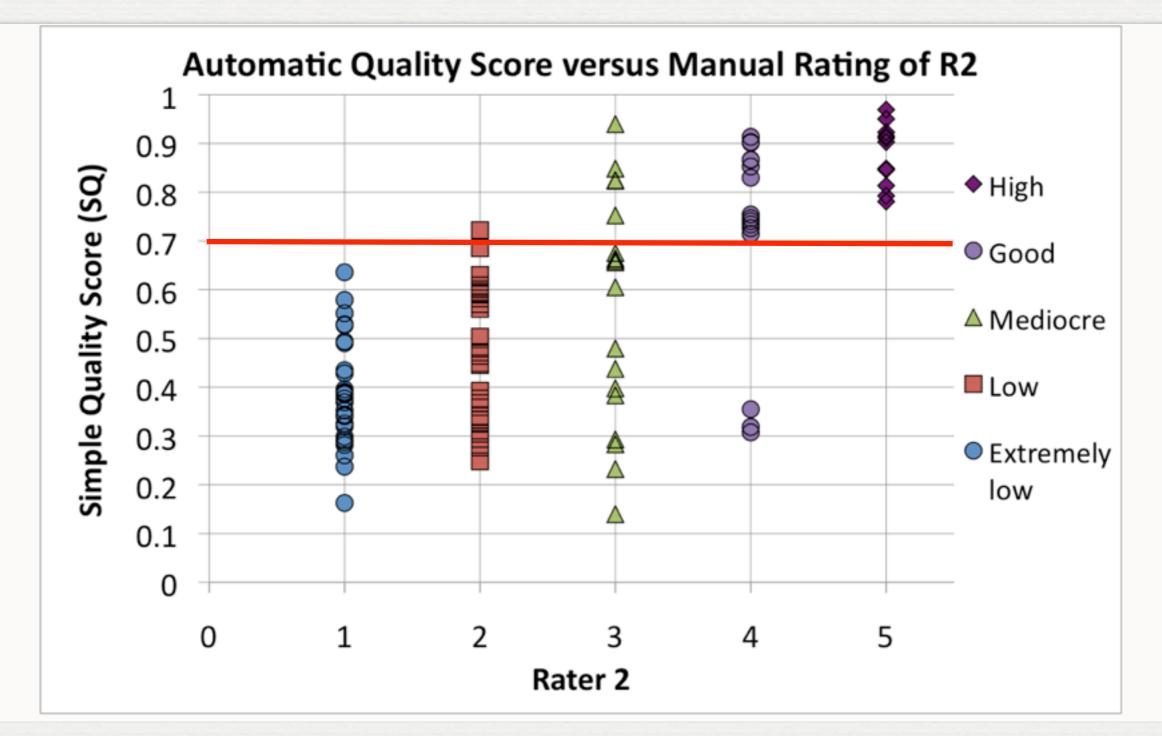
AUTOMATIC VS HUMAN



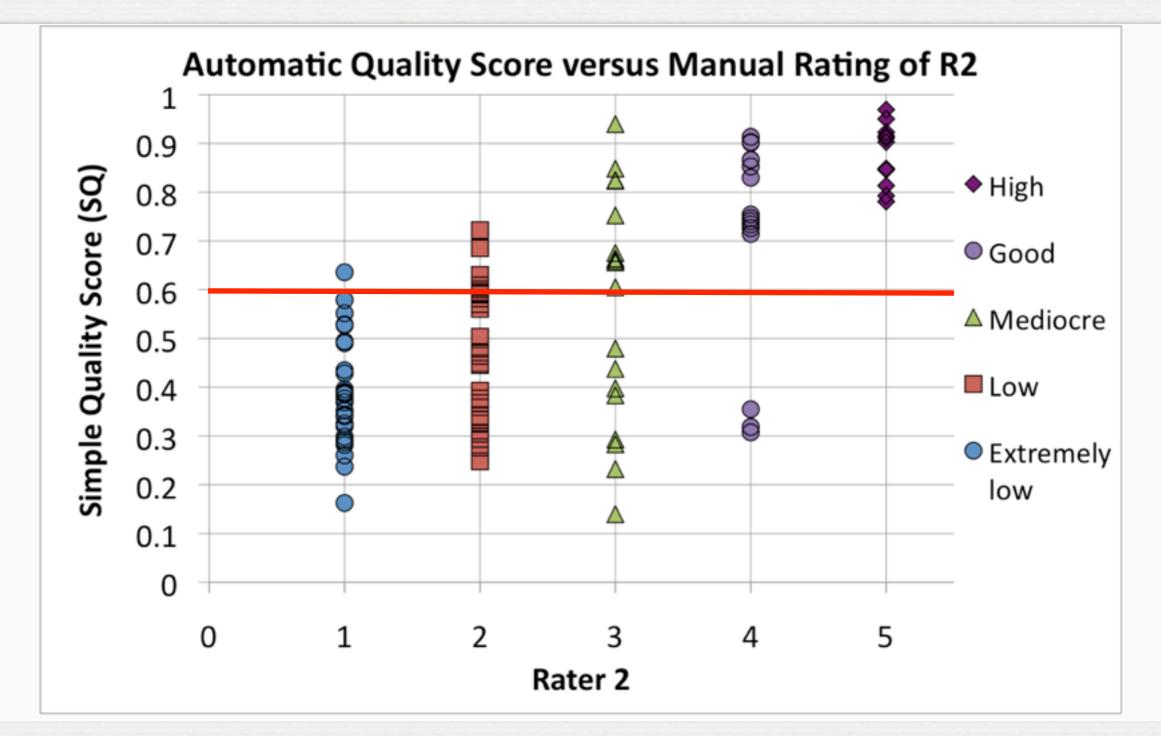
AUTOMATIC VS HUMAN



AUTOMATIC VS HUMAN

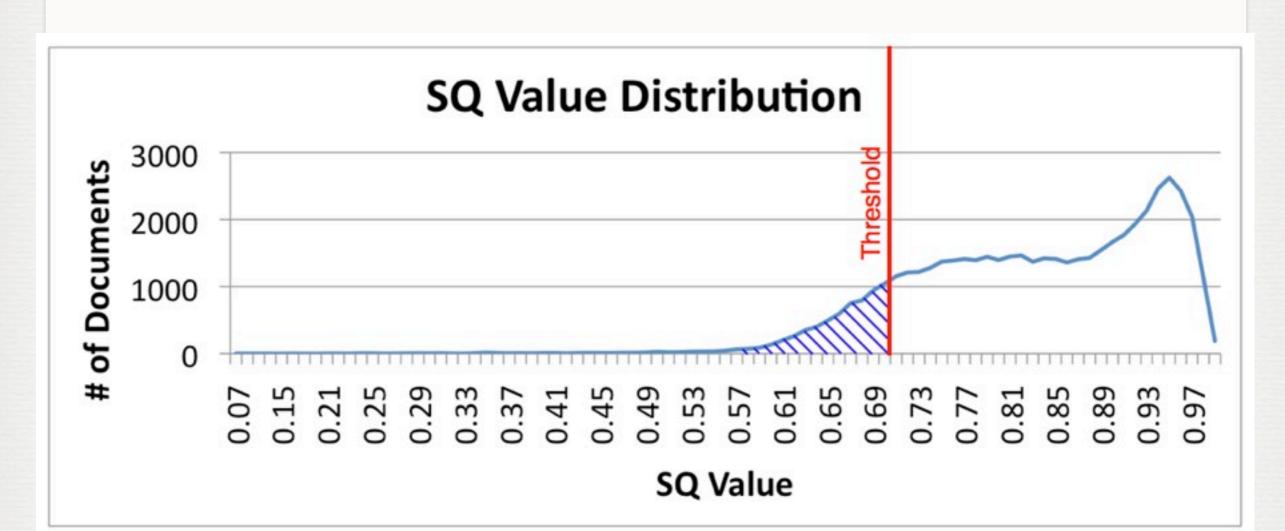


AUTOMATIC VS HUMAN



DATeCH 2014, May 20th 2014

THRESHOLD?



CONCLUSION

- We applied a simple quality scoring method to a large document collection and showed that automatic rating correlates with human rating.
- Document-level rating is not easy to do manually.
- Automatic document-level rating is not ideal but it give us a first "taste" of how good the quality of a document is. It is much more consistent than a person doing the same task.
- Many OCR errors are noise in big data but when added up they affect a significant amount of text.
- We found that named entities are effected worse than common words.
- HSS scholars need to be made much more aware of OCR errors affecting their search results for historical collections.

FUTURE WORK

- Consider publication date and digitisation date when doing OCR quality estimation.
- Examine the bad documents identify those worth post-correcting.
- AHRC big data project (Palimpsest) on mining and geo-referencing literature set in Edinburgh.
 Collaboration with literary scholars interested in locospecificity and its context in literature.

THANK YOU

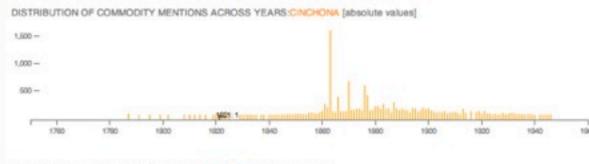


- Rating annotation guidelines and doubly rated data available on GitHub (digtrade)
- Contact: <u>balex@inf.ed.ac.uk</u>
- Website: <u>http://tradingconsequences.blogs.edina.ac.uk</u>/
- Twitter: @digtrade

BRINGING ARCHIVES ALIVE

ENTER A COMMODITY OF INTEREST Cinchona

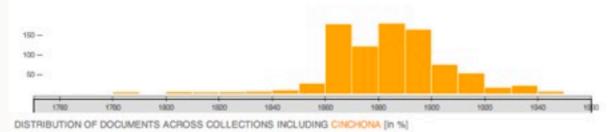
CINCHONA[®]



LOCATION MENTIONS ASSOCIATED WITH CINCHONA [absolute values]



DISTRIBUTION OF ALL DOCUMENTS BY DECADE [absolute values]



COMMODITIES RELATED TO Bark Camphor Cattle Cinnamomum Cassia Coal Cocoa Bean Coffee Copper Cotton Food Grain Fruit Gold Indigo Iron Lard Lead Lime Lumber Madras (Cloth) Maize Natural Rubber Opium Petroleum Pig Potash Quinine Rubia Tinctorum Sat Seed Silver (Color) Sugar Sutturic Acid Tea Tin Tobacco Vanilla Vegetable Vegetable lvory

DATeCH 2014, May 20th 2014



DOCUMENTS INCLUDING CINCHONA [top 100]

1863: East India (Chinchona Plant): Miscellaneous

1863: East India (chinchona plant). Return to an address of the Honourable the House of Commons, dated 9 March 1863;--for, "copy of correspondence relating to the introduction of the chinchona plant into India, and to proceedings connected with its cultivation, from March 1852 to March 1863."; House of Commons Parliamentary Papers

1870: East India (chinchona cutivation). Return to an address of the honourable the House of Commons, dated 3 May 1870;--for "copy of all correspondence between the Secretary of State for India and the Governor General, and the governors of Madras and Bombay, relating to the cutivation of chinchona plants, from April 1866 to April 1870."; House of Commons Parliamentary Papers

1876: East India (chinchona cultivation). Return to an address of the honourable the House of Commons, dated 8 July 1875;...for, copies of the chinchona correspondence (in continuation of return of 1870) from August 1870 k July 1875.; House of Commons Parliamentary Papers

1866: East India (chinchona plant). Return to an address of the Honourable the House of Commons, dated 14 May 1866;--for, "copy of further correspondence relating to the introduction of the chinchona plant into India, and to proceedings connected with its cultivation, from April 1863 to April 1866."; House of Commons Parliamentary Papers

1877: East India (Chinchona cutivation). Further return to an address of the honourable the House of Commons, dated 8 July 1875;--for, copies of the Chinchona correspondence (in continuation of return of 1870) from August 1870 to July 1875.; House of Commons Parliamentary Papers

1887: Ceylon in the "Jubilee year."; Miscellaneous

.....

■ 1883: Papers relating to Her Majesty's colonial possessions. Reports for 1882. (In continuation of [C.-3642.] of June 1883.); House of Commons Parliamentary Papers

 1885: Ceylon \& Her Planting Enterprize: In Tea, Cacao, Cardamoms, Cinchona, Coconut, and Areca Palms ...; Miscellaneous

1921: Trade and navigation. Return (in part) to an order of the Honourable the House of Commons, dated 16 February 1921;--for accounts relating to trade and navigation of the United Kingdom, for each month during the year 1921.; House of Commons Parliamentary Papers

In 1919: East India (Industrial Commission, 1916 18). Minutes of evidence taken before the Indian Industrial Commission, 1916-18. Vol. I.--Delhi, United Provinces, and Bihar and Orissa.; House of Commons Parliamentary Papers

1861: Notes on the medicinal Cinchona barks of New Granada by H. Karsten; and on the Cinchona trees of Huanuco (in Peru);

BRINGING ARCHIVES ALIVE

ABOUT

TE

enter a commodity of Interest
 cinchona
 raw data: cinchona

imum number of location mentions





enter a location of interest location search

top 50 location mentions by	/ decade									
1810	1820	1830	1840	1850	1350	1870	1880	1890	1900	1910
					Algeria	America	Africa	Aden	America	Amsterdiam
					America	Andes	America	America	Amsterdam	Argentina
					Andes	Asam	Assem	Amaterdam	Zatavia	Assem
					Andes	Assem	Avatralia	Asia	Sergel	Austria
					Australia	_ Autolia _	Sancroft	Australia	Sombay	Acengero
				America	Batavia	Bengal	Zalavia	Avatria	Casala	Zeiglum
				Andes	Bellary	Solvia	Bengal	Belgium	Ceylon	Sergel
				Andes	Bengal	Bombey	Rogola	Bengal	Chie	Suma
				Avatrala	Berhampore	Burna	Bombey	Canada	China	Calcutta
				Serin	Bolivia	Calcutta	Brazi	Casala	Derjeeing District	Carlabad
				Zoivia	Bombay	Castleton	Cacher	Ceylon	Europe	Ceylon
				Someo	Cachar	Ceylon	Calcutta	Chie	France	Channel Islands
				Brazi	Calcutta	China	Canada	China	Germany	China
				Zremen	Callout	Car.	Castleton	Colombia	Holand	Derjeeing
				Erunswick	Ceylon	Darjeeling	Ceylon	Derjeeling	Hong Kong	England
				Calfornia	China	East India	Crim	Derjeeing Datrict	Hutey	Europe
				Ceylon	Colmbatore District	England	Colombia	Egypt.		France
				Crime	Cuddapah	Europe	Colombo	England		Gambler
				Cleghorn	Cuenca	France	Darjeeling	Europe	India	Germany
				Cordillera	Darjeeling	Ganjam	Eg/01	France	Taily	Grate
				Deven	East India	Ganjam District	England	Germany	Javiaica	Great Britain
				East Inda	Ecuador	India	E-rope	Great Britain	Kelar	Gun
				Ecuador	Edinburgh		Gampola	Harris	Kelar District	Habue
				Edinburgh	England	Ineland	Germany	Hong Kong	La Merced	Hayo
				England	399 mentions	Jamaica	Great Britain		Lancashire	India
				Fakland Islands	Cusysquil	Kaahmir	India	India	Line	Tale Of Man
				Germany		Live paci		italy	London	Teaty
				Granada		London	Jamaica	Jamaica	Madras	Javaica
				Gualapita	in and	Madras	Java	London	Mager	Java
				Guayagul	isiay	Madrid	Kangra	Madras	Mauntica	Kasaul
				India	Jamaica	Malabar	Loridon	Matta	Montreal	Latore
				ina	Lima	Mexico	Madras	Mauntius	Mozembigue	Luxemburg
				Ireland	Loja	Nutra .	Mauritus	Mozembigue	Munge	Macassar
				Jamaica	Ldxs	Muttre	Michigan	Myacre	Netherlands	Madras
				Leiden	Madras	New York	Mysore	Nelherlands	Cr•	Maurtius
			Andes	Lind	Madura	Octacamund	New Zrunzwick	Nalgeri	Peru	Nelherlanda
			Drazi	Va	Mangalore	Cush	Nova Scotla	Ortano	Seycheles	Niger
			Comus	Markham	Maridham	Peru	Novara Eliya	Cr•	Shinoga	Oh
			Districts Of Montana	Numera Eliya	Masulipatam	Punjab	Ontario	Peru	Sierra Leone	Old Harbour
			England	Peru	Ootacamund	Russia	Oplacemund	Cuebec	South America	Cre
			Europe	Sectiand	Pegu Peru	Santa Fe	Peru	Rutate	Sgain	Punjab
			Fichida	South America		Sikhim South America	Punjab	Siam	St. Helens	Russia
			France	Spain	Quito		Quebec	South India	Targa	Spain
		America	Guayaquil	Sri Lanka	Rejehmundry District	Sri Lanka	South America Cel Lanka	Spater	The	Sumaira
		Litta	Montreal New Mexico	Stratford Swedien	South America	St. George St. Helens	Sri Lanka Triridad	Tunia	Tolu Trivided	Turkey United Kingdon
		Pavon	New Mexico Paris	Sweden Tierra Del Fuego	Southempton	St. Helens Timevell	United Kingdom	Turkey U.S.	Trinidad Turkey	United States
uiz De Fora		Payon	Para	Tierrs De Fuego Vries	Sri Lanka	Tinnevelly	United States	United Kingdom	United Kingdom	Vichy
Ourem	Bejuco	Rus	Pasco Peru	Walace	Tellichery	Tourgoo	West Africa	United States	United States	West Africa
irto De Mos					Trinidad					
NO DE MOS	Cordiers	Valgeralap	Tolu	Zambesi	10000	United States	West India	Zerober	Un	West Java
1810	1820	1830	1840	1850	1850	1870	1880	1890	1900	1910
alener .	1000	110000	1000	1000	1000	TOTO	1000	110000	1500	TEND.

BRINGING ARCHIVES ALIVE

Ø enter a commodity of interest cinchona raw data: cincho

uiz De Fora

Ourem

irto De Mos

Bejuco

Cordiers

enter a location of interest

ocation search

100 120 140 160 180 40 60 80

minimum number of location mentions





0 1840	1850	1860	1870	1880	1890
		Algeria	America	Africa	Aden
		America	Andes	America	America
		Andes	Asam	Assem	Amaterdam
		Andes	Assem	Avatrala	Asia
		Australia	Autole	Zancroft	Avatralia
	America	Batavla	Bengal	Zatavia	Austria
	Andes	Bellary	Bolvis	Bengal	Beiglum
	Andes	Bengal	Bombey	Rogola	Bengal
	Avatralia	Berhampore	Burra.	Bombey	Canada
	Zerin	Bolivia	Calcutta	Brazi	Casala
	Zoivis	Bombay	Castleton	Cacher	Ceylon
	Borneo	Cachar	Ceylon	Caloutta	Chie
	2mm	Calcutta	Crime	Canada	China
	Zremen	Callout	Car.	Castelon	Colombia
CINCHONA & INDIA [399 mentions in total]	Erunavick	Ceylon	Darjeeling	Ceylon	Detering
distribution of mentions (normalized by all commodity/k	ocation mentions from 1800-1920]	selected sentences from 186	0's documents follow to see full doc	ument list] Crime	Derjeeing District
0.20		1860 [] s sending Hooker :	some seeds of <u>Cinchona</u> [] <u>India</u> .		Eq.(*
0.20			nd seeds of <u>Cinchona</u> [] <u>India</u> on a	very large scale.	England
0.15		1860 [] Regarding the <u>Cir</u>	chona [] India.		Burge
			na [] India arrived, Mr Molvor found		France
0.10			on of the <u>Cinchona</u> [] <u>India</u> and Ce		Germany
			America with <u>Cinchona</u> [] <u>India</u> is e		Great Entain
0.05			related to <u>Cinchona</u> [] <u>India</u> and C	eylon [Srl Lanka].	Harris
			duction of Cinchona [] India		Hong Kong
1850 1870 1880	1890 1900 1910	Loss ([]) is a many mer and	the Cinchona [] India and gives h	is advice for theil	
	Fakaro Isanca	1861 [] posed journey to Ja	va to fetch Cinchona [_] india.		india
	Germany		Livergool	India	italy
	Grahada		London	Jamaica	Jamaica
	Guninguism	mana	Madras	20 m	London
	Guajagui	PIPI C			
			Madrid	Kangra	Madras
	India	Jamaica	Malabar	London	Mata
	Ins	Lima	Malabar Malabo	Lordon Medres	Mata Maurius
	ina Ineland	Lima Loja	Malabar Masico Muta	London Madras Mauritus	Maha Mautius Mazambiyue
	ina Indend Jamaica	Lima Loja Loxa	Malaban Malaba Nuta Muta	Lordon Madras Maurtus Mahgan	Mata Mauritus Noosmbigus Nyson
	ins Insind Jamaica Latan	Lima Loja Loka Madras	Malabar Mexico Multe Multe Naci	Lordon Madras Mautus Michgan Mysore	Mata Mauritus Mozarrólyjus Mydone Nacherlands
Andes	ina Indend Jamaica Lacian Lind	Lima Loja Loxa Madras Nadura	Master Meise Multe Multe New York Ootacamund	London Madres Meritus Michgan Nyton Nyton	Mata Mauritus Mozambiyue Mydone Nacherlands Nign
Zrazi	ins Insind Jamaica Ladan Lind Mai	Lima Loja Madras Madras Madras	Matter Mates Mutre Nation New York Otacimund Stan	London Madras Machtar Michan Mydan Mydan New Scala	Manta Maurtus Mastribus Mastrianda Nagri ⊐intano
Graci Corrus	ins Indend Jaminos Late Na Na Nati	Lima Loja Madras Nadura Nargalore Markham	Meter Meter Muter Nater Nater Nater Cotaspround Cota Cota Peru	London Madras Mavritus Margan Nyigan Nan Euronitak Nana Euronitak Nanasa Euronitak Nanasa Euronitak	Mata Maurtus Mozanbigue Nysore Natherianda Nign Crianto Crianto
Sreci Conus Datrota Of Nortana	ins Indend Jachica Laden Lind Mai Markham Nanata Elya	Lima Loja Madras Madura Mangalore Mangalore Masulpatam	Nextor Metro Metro Nextra NextYork Cotocomund Cotocomund Coto Peru Punjeo	Lordon Madras Neurtus Mahgan Nyisore Neu Suranick Neus Saate Nunari Siya Ortano	Mata Mauthus Masambayue Nyasre Natharanda Nign Ciniano Cinia Panu
Brest Comus Districts Of Montens England	ina Indend Jamica Latan Lata Nat Nat Nationa Nationa Nationa Nationa Nationa Nationa Nationa Nationa Nationa Nationa	Lima Loja Madras Madras Mangalore Markham Masulipatam Ootacamund	Nextor Nexto Nutra Nutra Nex York Ootacamund Outa Peru Peru Peru Putato Rusa	Lonton Madras Meritus Michgan Mystre New Spatis New Spatis New Spatis Charto Charto Charto	Nata Maufitus Notarrbinjue Native Native Native Native Native Native Contarro Cre Feru Cuebec
Dest Conus Datricts Of Montana England Europe	ina Indand Latian Latian Lati Nat Nati Natian Numeri Elya Panu Secland	Lima Loja Lois Madras Madras Masure Markham Masulpatam Ootacamund Pegu	Nester Neico Mutre Neiro Neiro Neiro Cotacismund Outen Peru Purjeo Rase Sarta Fe	Lonton Madras Martus Martus Myson Myson New Sects New Sects Numes Elys Conservue Conservue Seru	Mete Meurtus Mesenboya Nytone Netherance Nigos Crejano Cre Paru Cuetos Rutus
Breat Conva Datricta Of Montana England Europe Florida	ina Ineland Jamese Lacan Lac Nat Nation Nation Nation Nation Nation Nation Nation Nation Nation Sociand Socian Socian	Lima Loja Madras Nasiura Nangalore Narkham Nasulpatam Ootacamund Pegu Peru	Neter Meter Note Netry Netry Vote Coden Peru Ponjet Ruter Satis Fe Sikhim	London Madras Martus Mutan Nyson Nyson Nyson Nyson Socia Nyson Socia Nyson Socia Nyson Socia Disconund Penyso	Meta Maurikus Masarhöyue Nathvirane Nign Ordano Die Fany Cuebec Russa Siam
Eneri Coinus Datricta Of Ventana England Eluippe Finica France	ina Indiana Jaminas Jadan Laba Mai Mai Mai Mai Mai Mai Mai Mai Mai Ma	Lima Loja Lois Madras Nadura Nargalore Markham Masulipatam Ootacamund Pegu Peru Quito	Nexter Mexico Nutra Nextra Outocomund Cuton Peru Peru Peru Russa Santa Fe Sikhim South America	London Madras Maurtus Matrian Mytos New Socia Nove Socia Noves Sola Columno Columno Peru Punab Cuates	Mais Maurthus Mosambigue Nathairands Nign Creands Crea
Sest Conus Datricts Of Montane England Burlope Fights France tas Guity South	ina Indiano Jantinos Laden Laden Laden Markham Nuorata Etys Paru Sociano Socia	Lima Loja Lois Madras Madras Mangalore Masulipatam Ootacamund Pegu Pegu Guito Rajahmundiy District	Nester Metro Mutra Nestry Nest York Ootscamund Ootscamund Ootscamund Ootscamund South Peru Purpes Russ Sarta Fe Sikhim South America Sri Lanka	Lorison Madras Navrtus Martus Navra Navra Navra Sach Human Siya Orisos Cotacamund Sau Pungab Cuelac South America	Mets Mauffus Noteshboxe Natherands Nigri Creans Che Faru Cluebec Ruiss Siam South Inde Speier
Event Conva Datreta d'i Montana England Europe France France tea Guityjaquili Montanal	ina Indino Jaminos Ladan Ladan Undi Mai Narkam Narkam Narkam Narkam Sostand Sostand Sostand Sostand Sostand Sostand Sostand	Lima Loja Lois Madras Nasura Nasulgatore Naskham Nasulgatam Ootacamund Pegu Peru Quito Rajahmundiy District: Salam	Nester Neico Nota Neiro Neiro NeirYok Ootacamund Oota Peru Pupa Suna Sarta Fe Sikhim South America Sir Lanka Si. George	Lonton Madras Martus Michgan Myson New Sprawck New Sprawck New Sprawck New Sprawck Colasterund Delasterund Paru Puljab Colasterund Study America Study America	Main Maurtus Maurtus Migns Oriens Oriens Deu Peru Outeo: Ruiss Sem South Inda South Inda South Tuha
Brest Conus Datricts Of Montane Enrope Fisnos France res Guisjouil s Montreal on New Nacion	ina Inelect Jamese Lacan Lind Una Navere Bija Rev Sostend Sostend Sostend Sostend Sostend Sostend Sostend Sostend Sostend South America South America South America South Southea South America	Lima Loja Loia Madras Masura Margalore Markham Nasulgatam Ootacamund Pegu Peru Guito Rajahmudoy District South America	Nester Neico Nutre Nutre Nutre Nes York Ootschmund Outen Peru Purjeo Rasse Sarta Fe Sikhim South America Siri Lanka S. George S. Perue	Lonion Madras Martus Michan Myson Nyson Nyson Nyson Socia Nuson Scola Nuson Siya Chisan Colasamund Seru Punjab Culase South America South America Sin Lanka Triniad	Mein Maurtus Mashrbiya Nathrbinana Ngri Créana Créana Créana Créana San San San San San San San San San
Erest Conva Datricts Of Viertane Ergand Europe Finda France es Guty Soull a Montreal on Neir Vierto on Pairs	ina Ineland James Laban Und Markaam Narkaam Narkaam Narkaam Saaland Saaland Saaland Saalan Saaland Saalan Saalan Saaland Saalan Saaland Saaland Saaland	Lima Loja Lois Madras Nasura Nasulgatore Nasulgatore Nasulgatore Nasulgatore Nasulgatore Octacamund Pegu Peru Quito Rejetimundiy District Salem South America South America	Neter Netes Note Netry York Otach Otach Peru Purjao Rutan Sata Fe Silchim Bouth America Sri Lanka St. George St. Vena	Lordon Madras Martus Methan Nyato Nara Sorta Nyato Colacemund Colacemund Peru Punato Colacemund Peru Punato Colacemund Setu South America Bril Lanka Timicad United Orgotom	Mais Maurikus Masshibiyue Nathariandis Nigin Ordanto Dire Pinu Ouebac Russis Saim Stuth Indis Signifier Tunkay UIS.
Event Conva Datricta dr Nortana Erdjand Europe Finnis France nca Guitylocull a Montreal on Nava Vacio on Paria	ina Indiana Jantaica Jantaica Jantaica Jantaica Mai Markham Nuorata Eliya Paru Soland	Lima Loja Lois Madras Nadura Nargalore Narkham Nasulpatam Ootacamund Pegu Peru Quito Rajahmundiy District South America SouthAmerica Southamaton Si Lanka	Nester Melico Nutra Nesi York Ootscianrund Ootscianrund Oots Oota Peru Pojab Rutas Suta Suta Suta Suta Suta Suta Suta Suta	Lordon Madras Mauritus Matritus Matria Mysos New Socia Noves Socia Noves Socia Noves Socia Noves Socia Peru Purato Colesce Socia America Bri Lanka Trivisad United States	Mais Maurthus Maurthus Notore Nathelands Nign Creario
Even Conus Datricis dr Horises Ergand Burope Finns Prance ros Guojaquil e Montreal on New Nanco on Para u Parac	ina Indiana Jantaica Jantaica Laten Late Mai Markham Nuratra Eliya Paru South America South America South America South America South Sinaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Simaford Maisa	Lima Loja Loise Madras Nadura Nargalore Narkham Masulpatam Ootocamund Pegu Peru Quito Rajahmundiy Disolot South America Bouthampton Sri Lania Tellichery	Nester Metro Nutra Near York Cotastamund Cotastamund Cotas Peru Purpa Russa Santa Fe Sikhim South America Siri Lanka Si Ceorge Si, Heana Tinnevelly Tungso	Lordon Madras Mauritus Matrias Matrias Mars Soria New Soria New Soria New Soria Numers Sity Conserve Conserve Conserve Purge Conserve South America Sit Lanika Trinicad United States West Africa	Mais Maurtus Maurtus Notarends Nign Criento Crie Peru Cuebec Ruiss Sam South Inda Soeler Turkay US United Kingdom United Kingdom
Event Conva Datricta dr Nortana Erdjand Europe Finnis France nca Guitylocull a Montreal on Nava Vacio on Paria	ina Indiana Jantaica Jantaica Jantaica Jantaica Mai Markham Nuorata Eliya Paru Soland	Lima Loja Lois Madras Nadura Nargalore Narkham Nasulpatam Ootacamund Pegu Peru Quito Rajahmundiy District South America Bouthampton St Lanka	Nester Melico Nutra Nesi York Ootscianrund Ootscianrund Oots Oota Peru Pojab Rutas Suta Suta Suta Suta Suta Suta Suta Suta	Lordon Madras Mauritus Matritus Matria Mysos New Socia Noves Socia Noves Socia Noves Socia Noves Socia Peru Purato Colesce Socia America Bri Lanka Trivisad United States	Mais Maurthus Maurthus Notore Nathelands Nign Creario

DATeCH 2014, May 20th 2014

Amsterdam Derjeeling District rlong Kong Kolar District La Merced Lancashire Montreal Mozembique Netherlands Seycheles Shimoga Sierra Leone South America St. Helens United Kingdon

America

Zatavia

Sergel Sombay

Casala

Ceylon

Chie

China

Europe France

Germany Moland

Harley

÷.

÷. India

Italy

Jamaica

Kolar

Umo

London

Machas

Magini Mauritua

Mungo

One

Peru

Spain

Tanga The

тры

Trinicad

Turkey

United States

Ús.

Assem Austria Asseguro Regum Zengsi Sume Celevite Carlabad Ceylon Channel Islands China Derjeeing England Europe France Gambler Germany Grate Great Britain Guni Hague Hayti Incila Isle Of Man Italy. Jamaica Java Kasaul Lahore Luxemburg Macasan Madras Mauritua Netherlands Nigri dh Old Harbour Cre. Purjab Ratio Spain Sumatra Turkey United Kingdon United States Vichy West Africa West Java

ABOUT

21

Amsterdam

Argentine

Search by:

Commodity

Location



In Country GB Feature Type Capital Of Top-Level Administrative Division Population 435,791 GeoNames Entry View entry



SA

Documents in which 'Edinburgh' is mentioned in relation to commodities (Page 1 of 1)

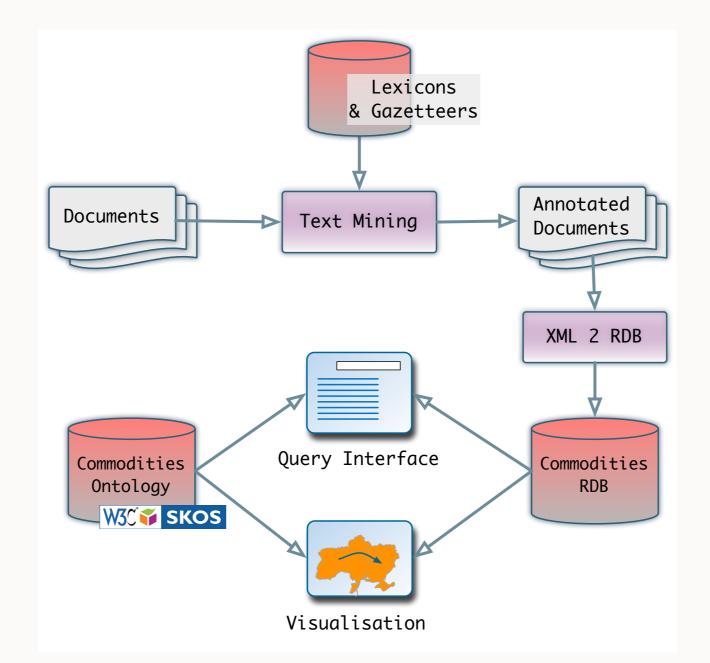
	Filtered by: 💼 Decade (1850) 🔡 Collection (House of Commons Parliamentary Papers) 💋 Commodity (Gold)						
16)	# Mentions	Document Title					
	90	Report from the Select Committee on the Bank Acts; together with the proceedings of the committee, minutes of evidence, appendix and index.					
	4	Twenty-ninth report of the Commissioners of Her Majesty's Woods, Forests and Land Revenues: in obedience to the acts of 10 George IV. (cap. 50), and 2 William IV. (cap. 1).					
	4	Parliamentary Papers. List of the bills, reports, estimates, and accounts and papers, printed by order of the House of Commons, and of the papers presented by command					

Filter

All House of Commons Parliamentary Papers (1 BY DECADE All 1850s (116) BY COMMODITY All Gold (116)

BY COLLECTION

SYSTEM



MINED INFORMATION

Example sentence:

From <mark>Padang</mark> was exported, in <mark>1871</mark>, <mark>6,127 piculs</mark> of cassia bark, of which a large portion was shipped to America (Fliickiger and Hanbury). ...

Normalised and grounded entities:

- commodity: cassia bark [concept: Cinnamomum cassia]
- date: 1871 (year=1871)
- Iocation: Padang (lat=-0.94924;long=100.35427;country=ID)
- Iocation: America (lat=39.76;long=-98.50;country=n/a)
- quantity + unit: 6,127 piculs

MINED INFORMATION

Example sentence:

From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Fliickiger and Hanbury). ...

Extracted entity attributes and relations:

- origin location: Padang
- destination location: America
- commodity-date relation: cassia bark 1871
- commodity–location relation: cassia bark Padang
- commodity–location relation: cassia bark America

EDINBURGH GEOPARSER



only 50 objects displayed, zoom in or deselect some features							
	Name	country	feature	km to center			
1 🖗	Ciudad Victoria 🏐	Mexico	seat of a first-order administrative division	11435.01 km			
2 🖗	Victoria 🏐	Seychelles	capital of a political entity	5126.1 km			
3 (P)	Victoria 🏐	Canada	seat of a first-order administrative division	11180.98 km			
4	State of Victoria 🏐	Australia	first-order administrative division	14782.65 km			
5 ®	Hong Kong 🏐	Hong Kong	capital of a political entity	9688.28 km			
<mark>6</mark> 🖗	Victoria 🏐	Malaysia	seat of a first-order administrative division	10558.79 km			
7 P	Durango 🏐	Mexico	seat of a first-order administrative division	11882.99 km			
8 🖗	Victoria 🏐	Malta	seat of a first-order administrative division	1494.57 km			
9	Victoria	Honduras	second-order administrative division	10832.96 km			
10 🖗	Victoria 🏐	United States	seat of a second-order administrative division	10954.12 km			



CONCLUSION

- Importance of two-way collaboration between technology and humanities expert in digital HSS projects.
- Value of iterative development and rapid prototyping.
- Geo-referencing text is very important for historical analysis.
- Most OCR errors are noise in big data but HSS scholars need to be made more aware of OCR errors affecting their search results for historical collections.