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



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



PROJECT OVERVIEW



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)

DOCUMENT COLLECTIONS

Co	llection	# of Documents	# of Images
House Parlia Early (Dire Corres Confide	Over 10 r Over 7	nillion docume 7 billion word t	ent pages, okens.
Asia and t	948,773 (OCRed: 450,841)		

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">
<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 refpectful 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,

alim8

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

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

Fast bugging her hehe to her bosom high swelling, Tight clasping her rosary, Mary's low bowed.

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

There allines a faint light like a star through a cloud; Upon the dark storm, from the fisherman's dwelling,

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.

Fishermen their vigils keep, Wherefore should their lovers weep?

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

OCR-ED TEXT

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 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;

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,

Watching the brooklets their pearly drops squander,

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

pap

<?>

<ar

< p

MR.

PRI

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	and 2000	omma-separated p	hrases: mohoo	jany,mabogany	with smoothing	of 3 🛟.		Q ⁺ Share y Twee	
.008 Search lots of b	ooks								
0.007									
0.000002%	mohogan y	/ E mabogar	y I						
0.0004 0.0000016%									
0.003 0.000012%									2
0.000008%									
0.0000004%				Λ	fl_			man	
0.00%	1680 1700	0 1720 1740	1760 1780	1800 1820	1840 1860	1880 190	0 1920 1	940 1960 1	980 2000

Google books Ngram Viewer

between 1730	and 2000 fr	om the corpus En	glish	with smoothing	of 🖪 🛟.	Twee	: {0
8 Search lots of b	ooks						
07							
06	mohogany	mabogany	mahogany				
0.0004%					1		
0.00032%					N	m	
0.00024%					\sim		
02				m	m	N	
0.00016%		M	mm				
0.0008%			m				s6
0.00%		- M					

Google books Ngram Viewer

een 1700 and 2000	from the corpus	English		with sm	oothing of	3 🗘 .	3	Tweet 0
rch lots of books								
time	lime							
0.15%								
0.40%	Jun	~~~~						
0.12%								
0.09%		-						
0.06%								
0.03%								
			257					

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













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

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

0 20 40 60 80 100 120 140 160 180

imum number of location mentions





enter a location of interest location search

1810	1820	1830	1840	1850	1850	1870	1880	1890	1900	1910
					Algeria	America	Africa	Aden	America	Amsterdam
					America	Andes	America	America	Amsterdam	Argentina
					Andes	Assem	Assem	Amaterdam	Zafavia	Assem
					Andes	Assem	Australia	Asta	Sergel	Austria
					Australia	Australia	Zancroft.	Australia	Zombay	Asengero
				America	Batavla	Bengal	Zafavia	Austria	Casala	Zeigium
				Andes	Bellary	Solvia	Bengal	Belgium	Ceylon	Sergs
				Andes	Bengal	Bombay	Zogola	Bengal	C10	Burma
				Australia	Berhampore	Burris .	Bombey	Canada	China	Calcutte
				Zerin	Bolivia	Calcutta	2reci	Casala	Derjeeling District	Caristed
				Zolvis	Bombey	Castleton	Cachar	Ceylon	Europe	Ceylon
				Someo	Cachar	Ceylon	Caloutta	C+++	France	Channel Islands
				Stati	Calcutta	China	Canada	Chine	Germany	China
				Zremen	Cellout	Cer.	Castleion	Colombia	Holand	Corjecting
				Erunpisick	Ceylon	Darjeeling	Ceylon	Derjeeling	Hong Kong	England
				California	China	East India	Chris	Darjeeing District	Hurley	Europe
				Ceylon	Colmbatore District	England	Colombia	Egypt.		France
				China	Cuddapah	Europe	Colombo	England		Gambler
				Ceptorn	Cuenca	France	Darjeeling	Europe	India	Germany
				Cordillera	Darjeeling	Ganjam	Egypt.	France	Italy	Grate
				Detren	East India	Genjem District	England	Germany	Jamaica	Great Britain
				East Inde	Ecuador	Los all as	Europe	Great Britain	Kolar	Gun
				Ecuador	Edinburgh	india	Gampola	Parts	Kelar Datrict	Hague
				Edinburgh	England	Ireland	Germany	Hong Kong	La Merced	Hayti
				England	Europe	Jamaica	Great Britain		Lancashire	Incila
				Fakland Islands	Gueyaguil	Kashmir	d	India	Line	Isle Of Man
				Germany		Livergool	India	Italy	London	Italy
				Granada	India	London	Jamaica	Jamaica	Madras	Jamaica
				Gualaputas	IIIula	Madras	and a	London	Magiri	an an
				Gueyecul	15(5)	Madrid	Kangra	Madras	Mauritus	Kanaul
				India	Jamaica	Malabar	London	Mata	Montreal	Lahore
				ina.	Lima	Mexico	Madras	Mauritus	Mozembique	Luxemburg
				Ireland	Loja	Muta.	Mauritus	Mozambique	Munga	Macanaar
				Jamaica	Loxa	Matra	Michigan	Mysore	Netherlands	Madras
				Leiden	Madras	New York	Myapre	Netherlands	0.0	Mauritus
			Andes	Line	Madura	Octacamund	New Enumenick	Nen	Peru	Netherlands
			2 maril		Mandalore	Puth	Nova Scotia	Ortario	Severheim	Nex
			Corpus	Markham	Markham	Peru	Name Ele	-	Shimona	-
			Detricts Of Mexicos	Name Elect	Masulloatam	P. cm	Detatio	Part.	Siarra Lanna	Cid Partner
			Enderd	Peru	Optacamund	P. ania	Delacamund	Cusher	South America	
			Europe	Sector	Pequ	Same Fe	Per	P.ana	South	Punish
			Elevera	South America	Peru	Sikhim	P. cmb	5	S. Halana	P. esta
			Erenza	S-an	Ruito	South America	Distant.	So the local	Terre	Seato
		Arrest a	Guyenul	Set Lanks	Balahmundry District	Sri Lanka	South America	State	The	Sumaira
		Line	Montreal	Similard	Salem	St. Germe	Sri Lanka	Turia	Telu	Turkey
			New Mexice.	Sweden	South America	St. Halana	Transa	Turkey	Transfer	United Kinodon
			Date	Tierre Del Suerro	Southernotice	Trendel	Linker Contem	LUS S	Turker	United States
uiz De Fora		- North	Para	View Proventier	Srillanka	Tinnevelly	United States	Lipited Kingdom	Lipited Kingdom	Vite
Ourem	Ee uro		- Sec.	147	Tellichery	T	West Africa	Linited States	Linited States	West Street
into De Mos					Tublicat					THE PROPERTY AND A DESCRIPTION OF THE PR
NO DE NOS	Coreiters	A B C B B B C B C B C B B C	100	2.877.048	1111220	United Science	VVAR PICK	2 Bracer	Ca.	11 M M M
1810	1820	1880	1840	1850	1850	1870	10000	1830	1900	1910

BRINGING ARCHIVES ALIVE

Ø enter a commodity of interest cinchona raw data: cincho

uiz De Fora

Ourem

rto De Mos

cation search

minimum number of location mentions

80

100 120 140 160 180

enter a location of interest

top 50 location mentions by decade

1820	1830	1840	1850	1860	1870	1880	1890	1900
				Algeria	America	Africa	Aden	America
				America	Andes	America	America	Amsterdam
				Andes	Assem	Assem	Amaterdam	Satavia
				Andes	Assem	Australia	Asis	Zengsi
				Australia	Autoin Donacol	Eancroft	Australia	Sombay
			America	Batavla	Bengai	Zafavia	Austria	Cassia
			Andes	Bellary	Bolivis	Bengal	Belglum	Ceylon
			Andes	Bengal	Bombay	Zogola	Bengal	Chie
			Australia	Berhämpore	2.uma	Bombey	Canada	China
			Derin	Bolivia	Calcutta	Drazi	Cassia	Darjeeing Datri
			Zolvis	Bombay	Castieion	Cachar	Ceylon	Europe
			Eomeo	Calcutta	Ceylon	Calcutta	C10	France
			2/mpl	Calculta	Crime	Canada	China	Germany
			Bremen	Callout	Der	Castielon	Colombia	Moland
	CINCHONA	& INDIA [399 mentions in total]	Zrumawick		Darjeeling		Deteeing	Hong Kong
	distribution of	of mentions [normalized by all commodity/	location mentions from-1800-1920]	selected sentences from	1860's documents [click to see full do	cument list] Crime	Dejeeing Datrict	Hurley
	0.20			1860 [] s sending Hooks	er some seeds of <u>Cinchona</u> [] <u>india</u> .		=c);0 ⁴	
				1860 [] procuring Plants	s and seeds of <u>Cinchona</u> [] <u>india</u> on a	a very large scale.	England	
	0.15			1860 [] Regarding the g	<u>Cinchona</u> [] <u>India</u> .		Europe	Inclia
				1861 [] When the [Cinci	nona [] India arrived, Mr Mcivor foun	o all	France	Italy
	0.10			1861 [] ed with the cultiv	vation of the <u>cinchona</u> [] <u>India</u> and Ce	eyion, contains the resu	Germany	Jamaica
	0.05			1861 [] ignee from Sou	in America win <u>Cinchona () India</u> is ets related to Cinchona () India and C	expected to arrive at Call Sevice (Scill Seks)	Great 2rtain	Kolar
				1861 [] Bombay on the k	ntroduction of Cinchons [1] India	eyion (on cankaj.	Carra -	Koar Dane.
	0.00		In	- It list i _ 1861 [_] Edition of the	nce the Cinchona [_] india and gives h	is advice for thei	Hong Kong	La Merced
		1850 1870 1880	1890 1900 1910	1861 [] posed lourney to	Java to fetch Cinchona [_] India			Lancashire
			- BKBYC I BBYCS			India		Lime
			Germany	India	Lowgood	lantaica	italy	London
			Contractor Contractor		Madras	Jamaica	London	Nacras -
			Genegoen		Madda		Madaz	Nager -
					10000		1000	Verteel
			-	Lima	Marine	Madras	March a	Maramhia
			ing and		14.44		Manager Manager	1. And a state of the state of
				Lova	10.41	Michigan	Marrie	Netherlands
			Leter	Madras	New York	Muser•	Nelberianda	Gr.
		Andes	Lind	Medure	Octacamund	New Stumpick	Nist	Peru
		2 mm		Mangalore	Port.	Nova Scola	Column	Sauchalas
		Carnus	Nackbarr	Markham	Peru	Numera Elica	S .	Shimona
		Districts Of Montana	Notes Eles	Masulloatam	Punish	Ontario	Peru	Sierra Leone
		England	Peru	Octacamund	Russia	Optingemund	Quebec	South America
		Europe	Scolland	Pegu	Santa Fe	Peru	Russia	Spain
		Florida	South America	Peru	Sikhim	Punjab	Siam	St. Helens
		France	Spain	Quito	South America	Cuebec	South Inde	Tanga
	America	Guayaguli	Sri Lanka	Rajahmundry District	Sri Lanka	South America	Spelar	The
	Uma .	Montreal	Stratford	Salem	St. George	Sri Lanka	Tunix	Telu
	London	New Vexico	Sweden	South America	St. Helens	Trinidad	Turkey	Trinidad
	Payon	Para	Tierra Del Ruego	Southempton	Timevell	United Kingdom	u s.	Turkey
	Peru	Pasco	Vries	Sri Lanka	Tinnevelly	United States	United Kingdom	United Kingdo
Bejuco	Rus	Peru	Walace	Tellicheny	Toungeo	West Africa	United States	United States
Cordilers	Valgerates	Telu	Zambeal	Trinidad	United States	West India	Zarobar	Un Un
1820	1830	1840	1850	1860	1870	1880	1890	1900

DATeCH 2014, May 20th 2014

Assem Austria Acargaro Segum Zengal Sume Calcutta Carlabad Ceylon Channel Islands China Derjeeling England Europe France Gambler Germany Grate Great Entail Guni Hagua Hayti India Isle Of Mar Italy Jamaica Java Kasaul Lahore Lucemburg Vacanar Macros Maurtius Netherlanda Nigri dh Old Harbou Cre. Punjab Ratio Scein Sumatra Turkey United Kingdon United States Vichy.

West Africa

West Java

DE

Amsterdam Argentine

ABOUT

Search by:

Commodity

Location



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



SA

Filter

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

	Filtered by:	Decade (1850) House of Commons Parliamentary Papers) 🖉 Commodity (Gold)
6)	# 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

All
House of Commons Parliamentary Papers (11
BY DECADE
1850s (116)
BY COMMODITY
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 🖗	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						
6 (P)	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.