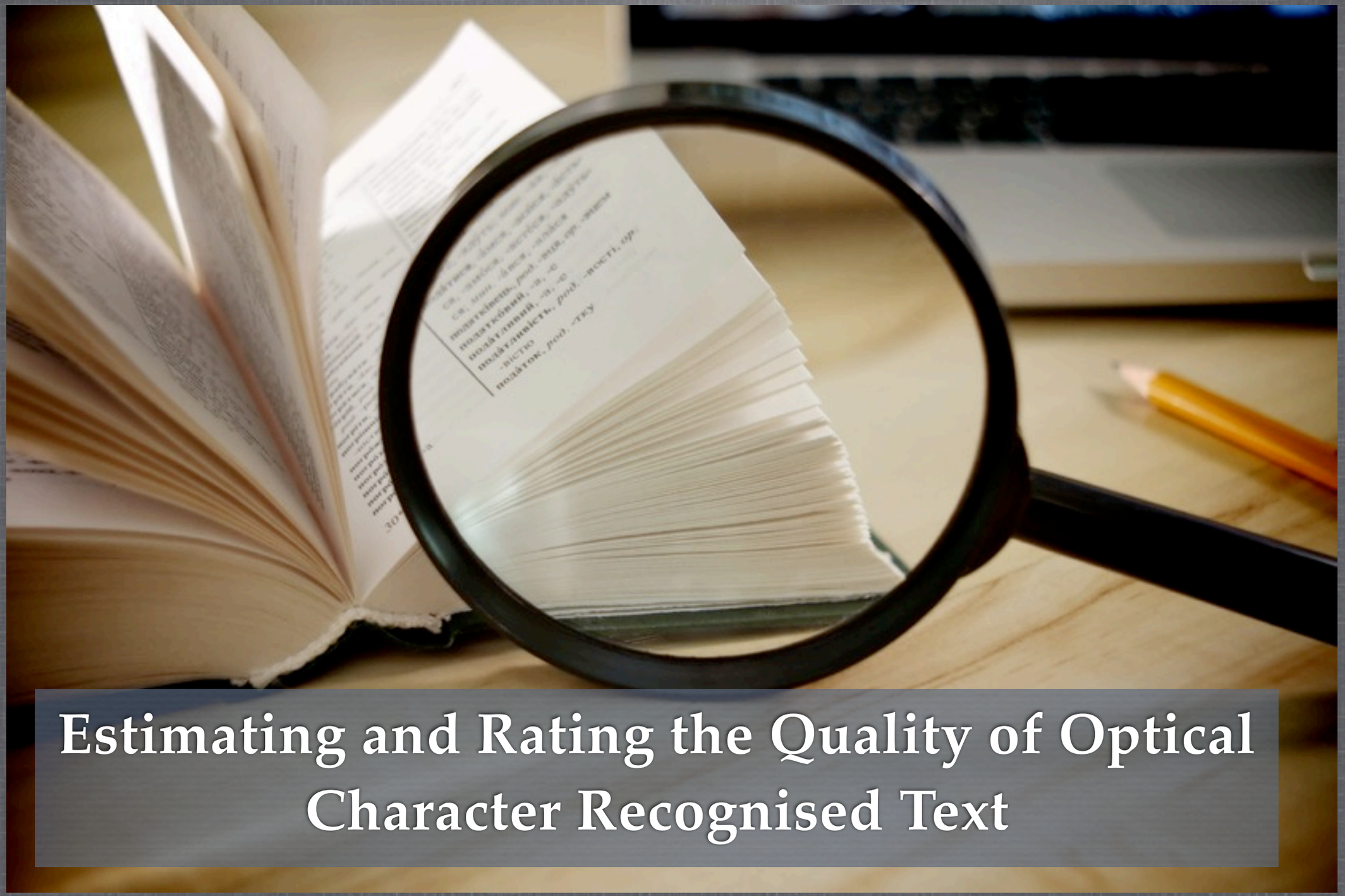


Estimating and Rating the Quality of Optical Character Recognised Text

THE UNIVERSITY of EDINBURGH informatics

Beatrice Alex

balex@inf.ed.ac.uk

Trading Consequences

DATeCH 2014, May 20th 2014

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*

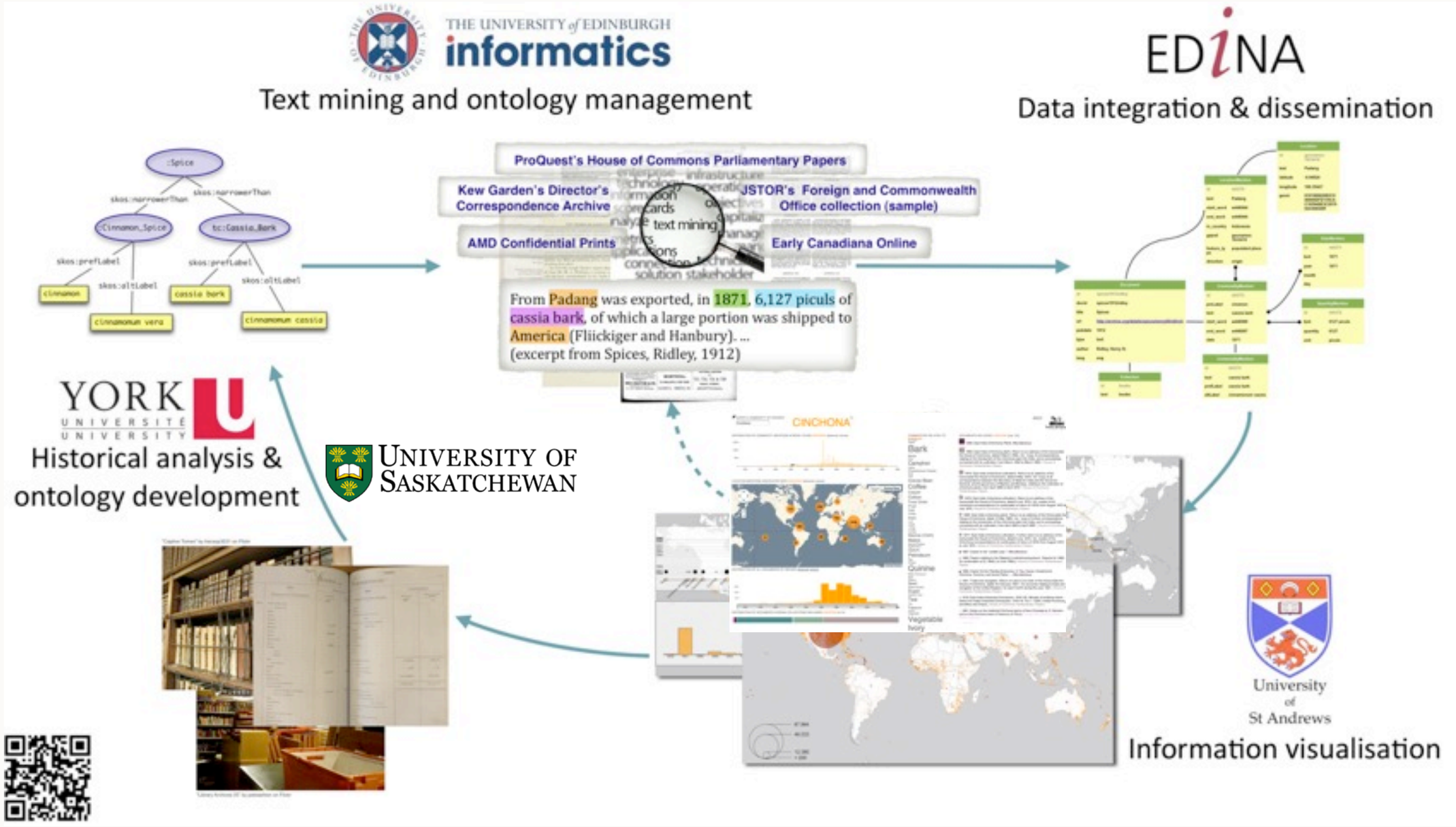
TRADITIONAL HISTORICAL RESEARCH

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.

Global Fats Supply 1894-98

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)

DOCUMENT COLLECTIONS

Collection	# of Documents	# of Images
House of Commons Parliamentary Papers		
Early C		
Dire Corres		
Confide		
Comm		
Asia and the West (Gale)	4,725	948,773 (OCRed: 450,841)

Over 10 million document pages,
Over 7 billion word tokens.

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>
<page> <![CDATA[T, G1IXET, Printer.]]> </page>
<page> <![CDATA[' INTRODUCTION, AS the following Vindication may fall into the hands of perfons who have
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>
<page> <![CDATA[4 2. Toward the clofe of the fame year the valuable 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 Na¬
tional 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 refpeclful 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">
 <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>
```

```
Proclamations, Pro * v mie RL’ E.LI S B.AIGO7.
iVICTORlfIA. h&gt;I l/ t(aŤ’’ of’ GO!&gt;. tif ih Firi.
ea fil~~Ť/ r&lt;’ lluil’tIŤ, (i’. i’ , QUEE’. Tc ,iii-
n 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 &lt; 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 refpeclful 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">
 <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>
```

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

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

She looks through her tears to her Father in heaven,
Then, kissing her child, cries, "Ah, bless'd be that smile!
And bless'd be the saints! now the promise is given!
Sure, mass it was said—he'll he here in a while."

Again and again with fond kisses bedewing
Her baby's soft cheek, as the rosebud as fair;
Again and again her look upward renewing,
This song, sweet and solemn, she breathes on the air:

When upon the stormy deep,
Fishermen their vigils keep,
Wherefore should their lovers weep?
Thou art there.
When the storm is raging high,
And the waves leap to the sky,
Wherefore should we fret and sigh?
Thou dost care.

Stronger than the tempest's rage
Is Thy power from age to age;
Thou didst once its wrath assuage,
And can still.
We will own Thy mighty power,
Even in the darkest hour,
And receive as blessed dower
All Thy will.

The Fishermen's Mass.

By fair Tralee bay, oh! did you e'er wander,
Where quiet and beauty are blended in one,
Watching the brooklets their pearly drops squander,
Bright gems soon to form for the brow of the sun?
And have you e'er heard the swell of devotion,
Like chorus of angels from depths of the sea,
As mass hath been said by those sons of the ocean,
The fishermen humble of quiet Tralee?

Oh! I have been there, and seen them all kneeling,
Their faces turned upward beneath the bright sky,
And heard their low throbs of worshipping feeling,
And seen pure devotion gush out from each eye;
And oft, 'mid the worlds rough battles and sorrow,
I've thought of that fishermen's mass of the sea,
And pray'd for such faith in brighter to-morrow
As shone from each face in the bay of Tralee.

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

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

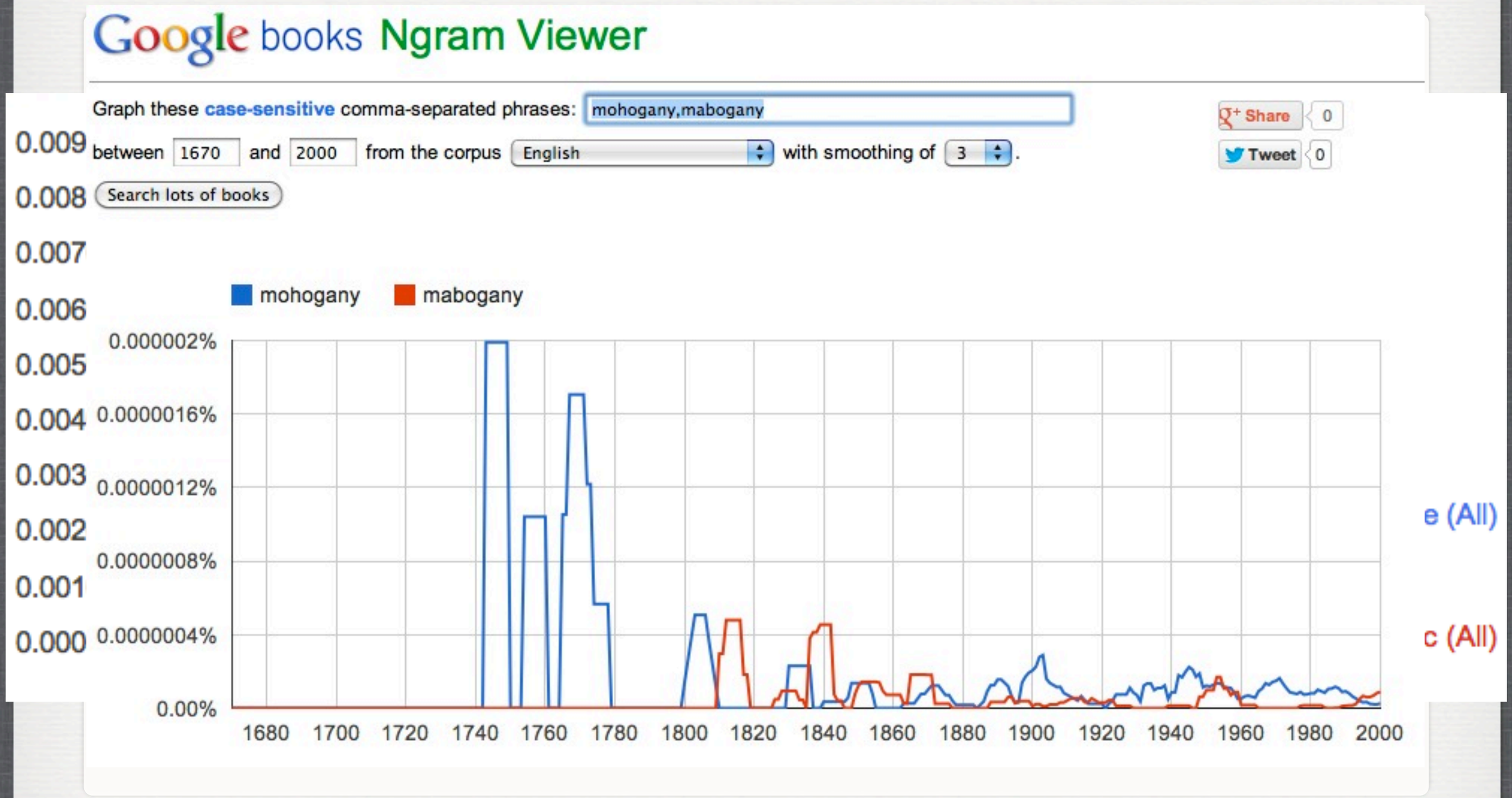
OCR ERRORS


0.000000600%
0.000000550%
0.000000500%
0.000000450%
0.000000400%
0.000000350%
0.000000300%
0.000000250%
0.000000200%
0.000000150%
0.000000100%
0.000000050%
0.000000000%
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
cattlc
Cattlc

OCR ERRORS

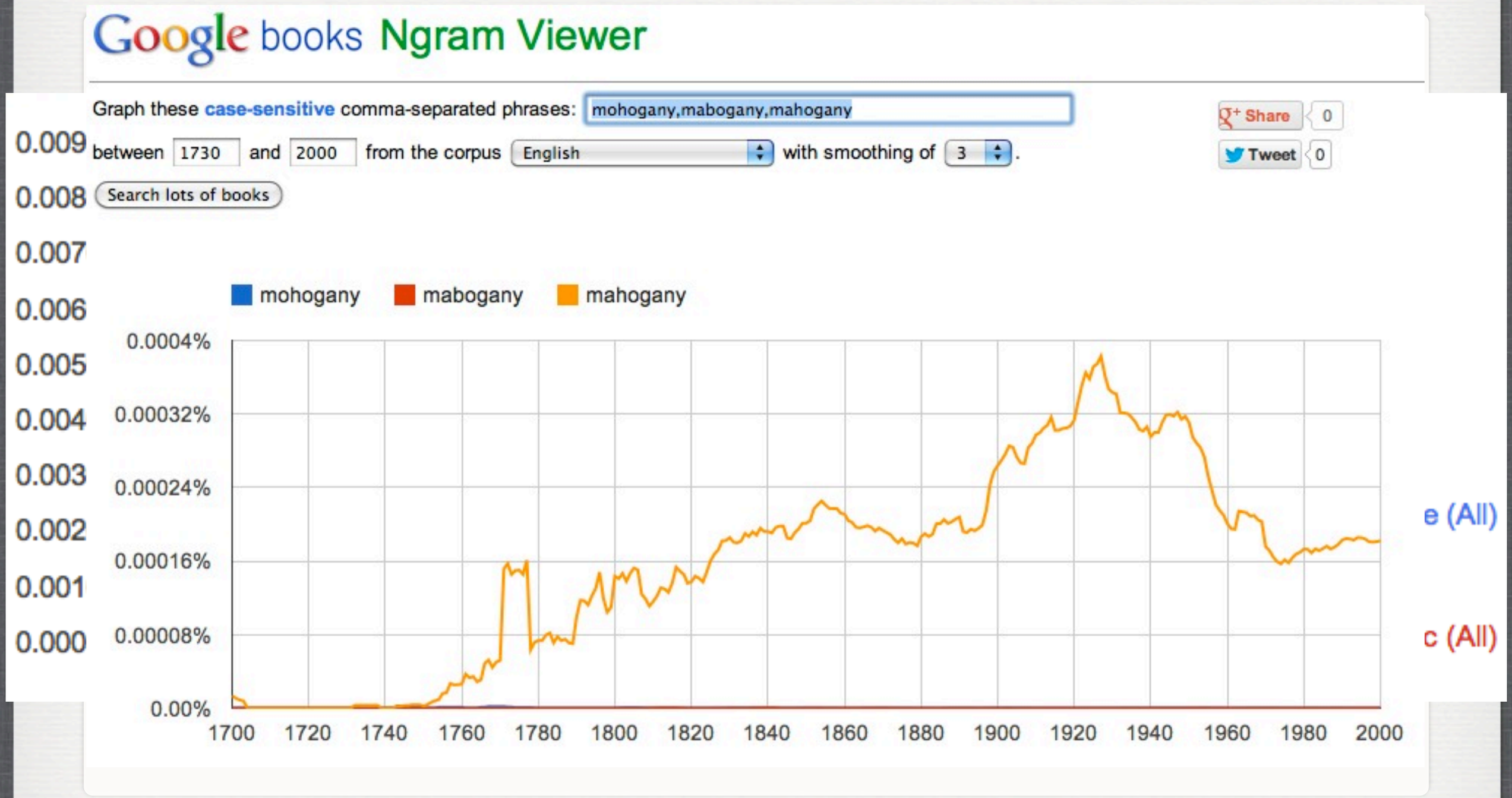

0.00900%
0.00800%
0.00700%
0.00600%
0.00500%
0.00400%
0.00300%
0.00200%
0.00100%
0.00000%
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
cattle (All)
cattlc (All)

OCR ERRORS

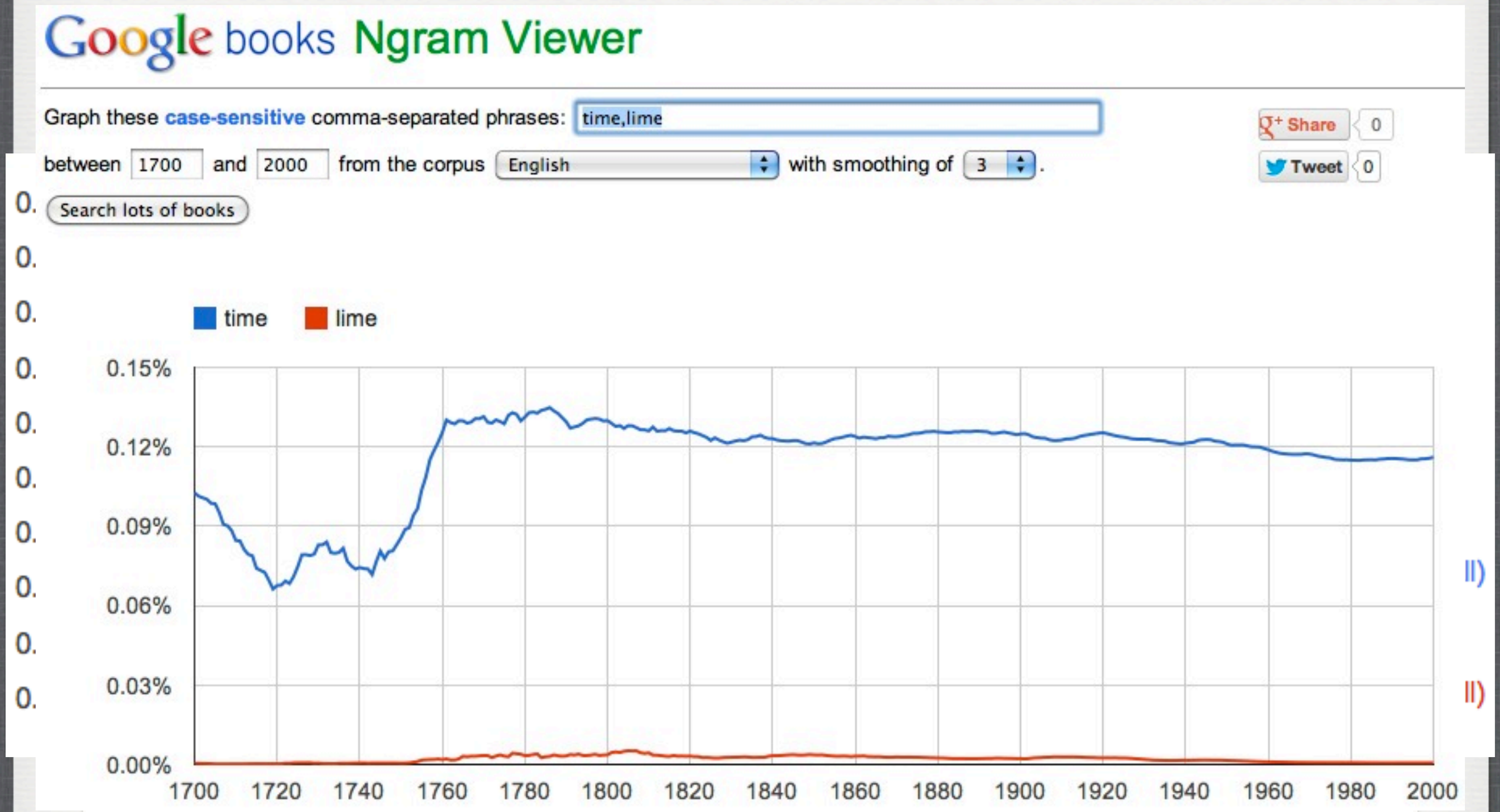


PQIS All Team Meeting, ProQuest, April 23rd 2014

OCR ERRORS


Google books Ngram Viewer
Graph these case-sensitive comma-separated phrases: mohogany,mabogany,mahogany
between 1730 and 2000 from the corpus English with smoothing of 3 .
Search lots of books
Share 0
Tweet 0
mohogany
mabogany
mahogany
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
0.0004%
0.00032%
0.00024%
0.00016%
0.00008%
0.00%
1700 1720 1740 1760 1780 1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000
e (All)
c (All)

OCR ERRORS



PQIS All Team Meeting, ProQuest, April 23rd 2014

OCR ERRORS AND TEXT MINING

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

	TP	FP	FN	P	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?

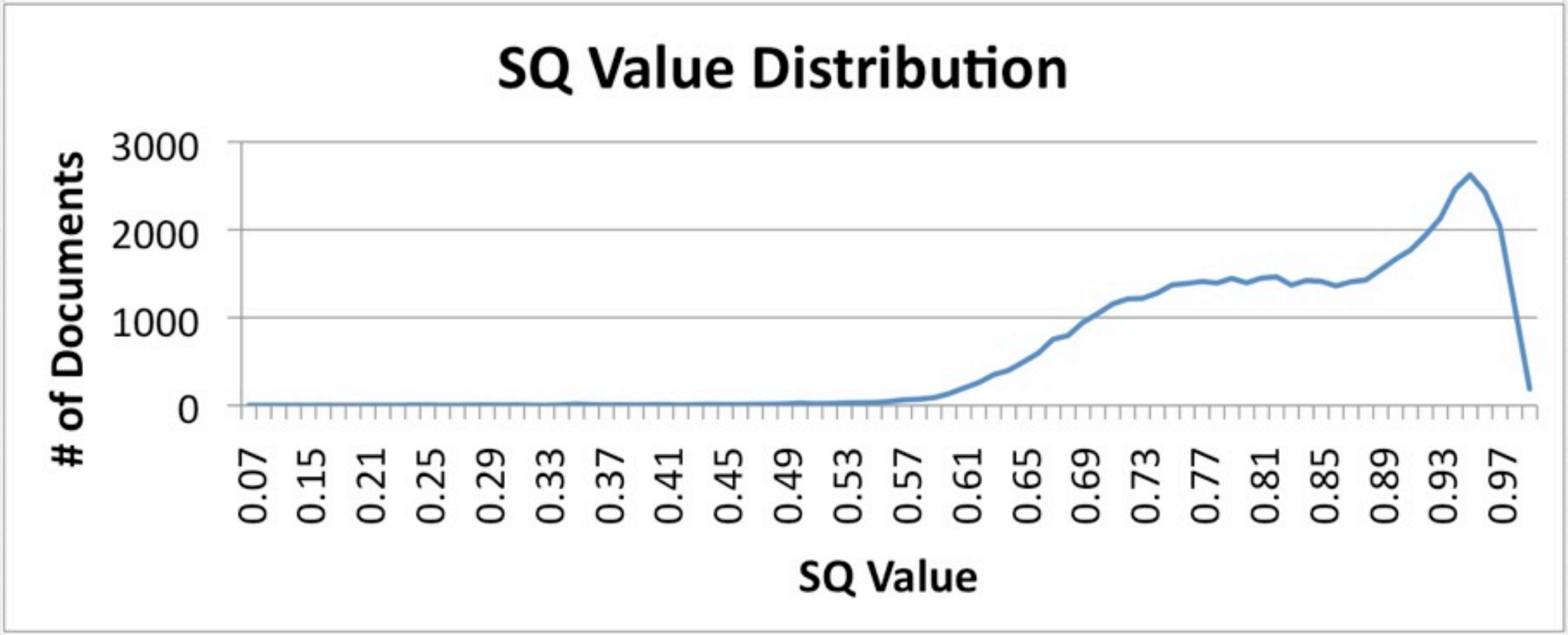
COMPUTING TEXT QUALITY

- 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 >= SQ < 0:2,
  - 20 documents where 0.2 >= SQ < 0.4,
  - 20 documents where 0.4 >= SQ < 0.6,
  - 20 documents where 0.6 >= SQ < 0.8,
  - 20 documents where 0.8 >= 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.

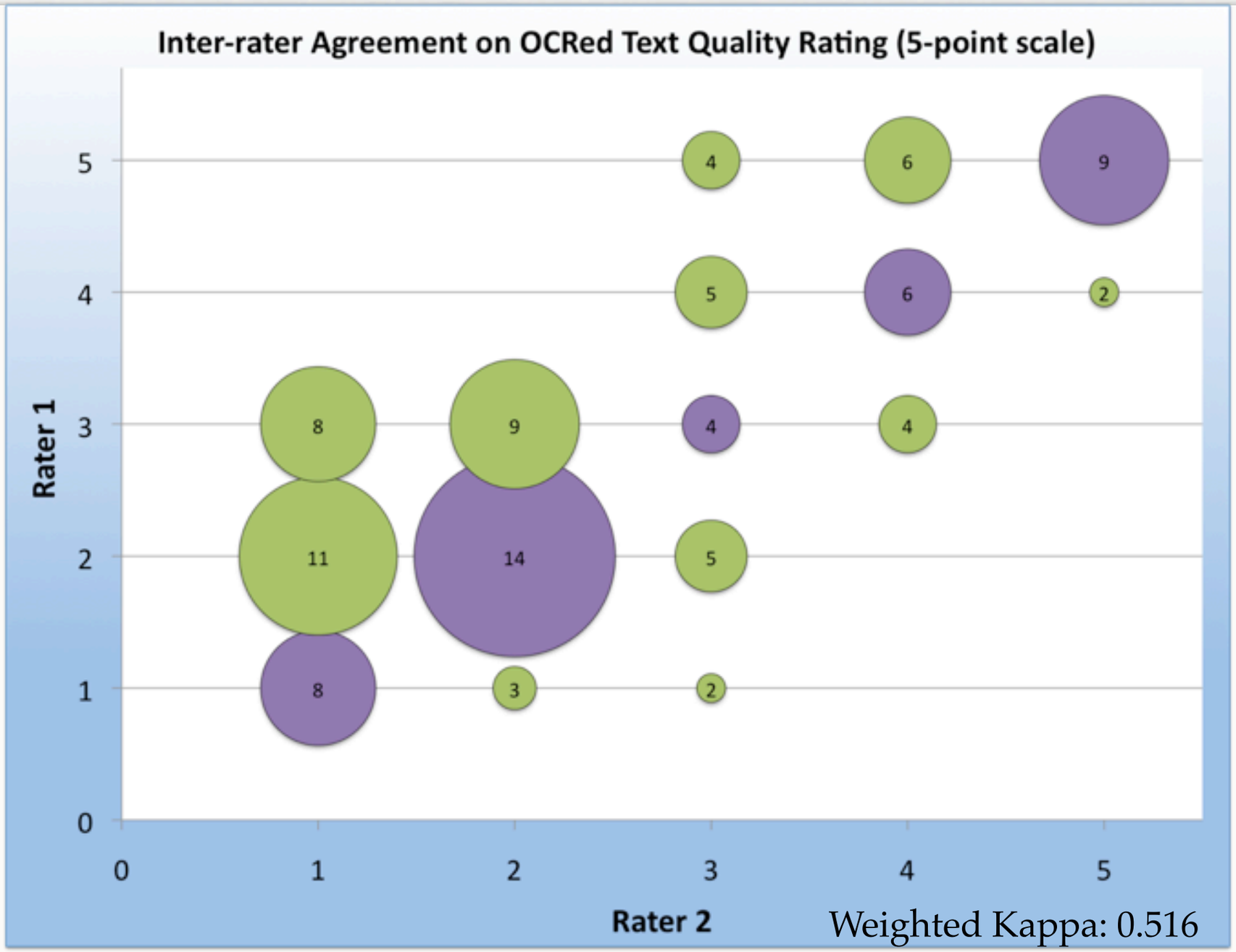
  4 ... OCR quality is good. There are some errors but they are limited in number and the text is still mostly readable and understandable.

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

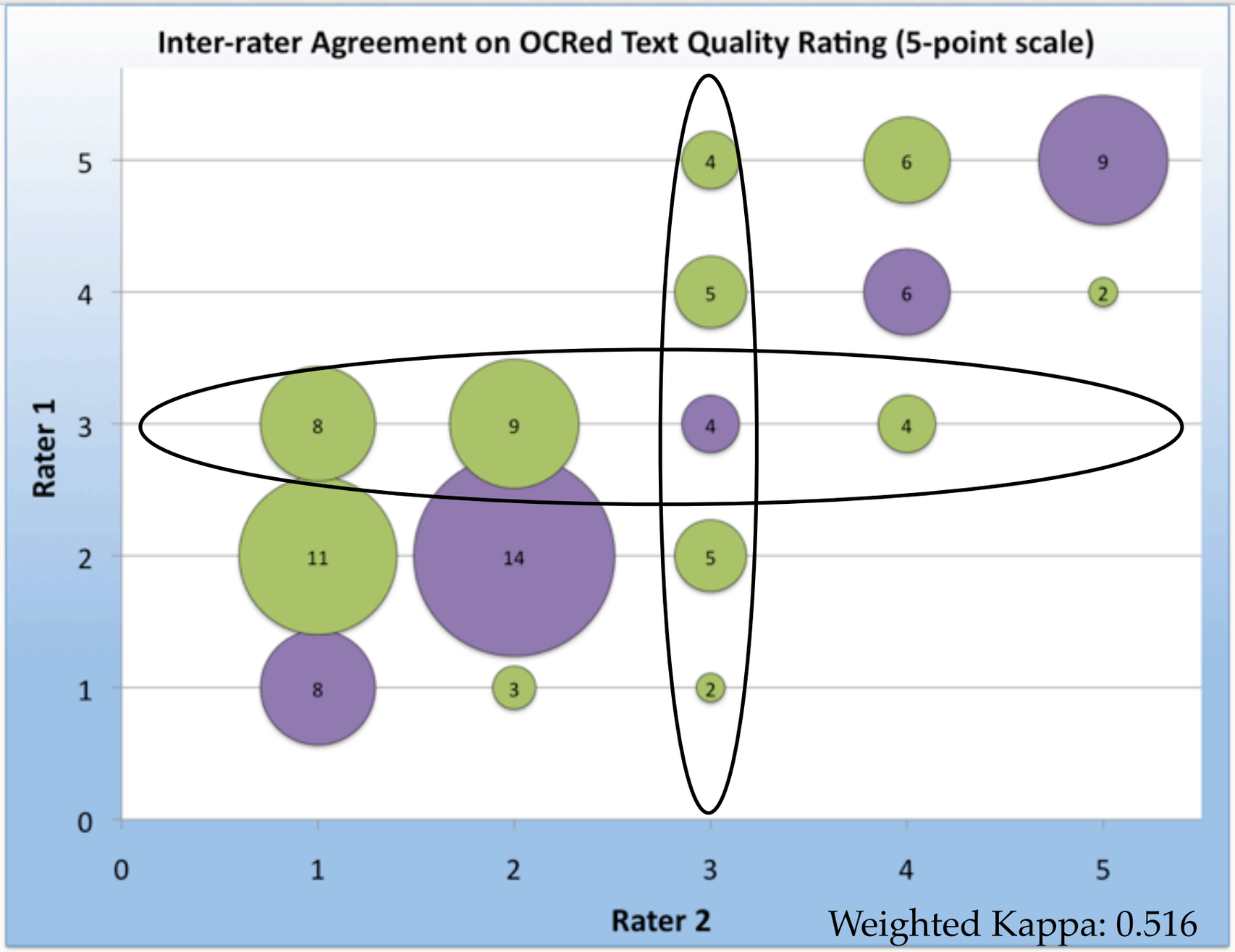
  2 ... OCR quality is low. There is a large number of OCR errors which seriously affect the readability and understandability of the majority of the text.

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

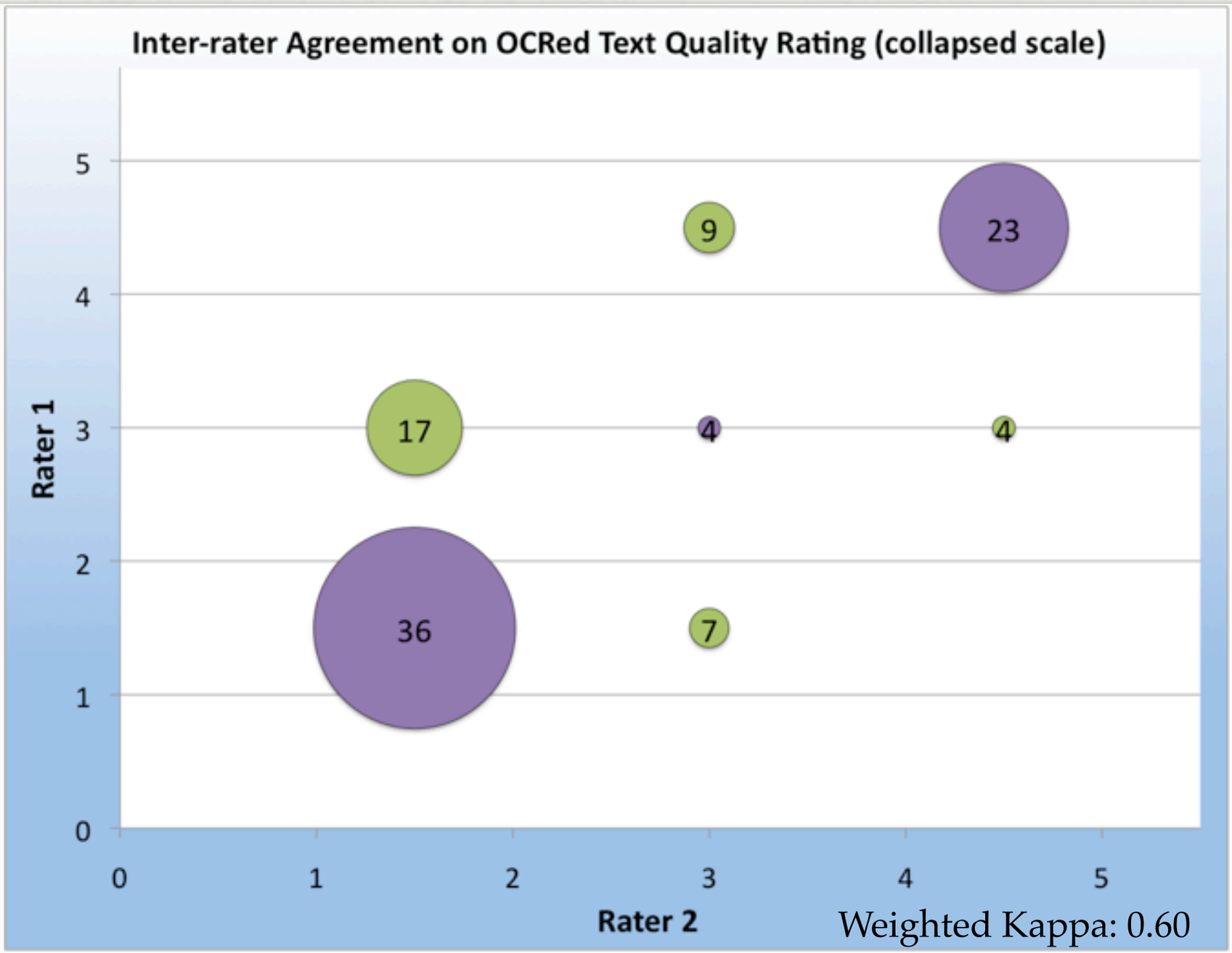
INTER-RATER AGREEMENT



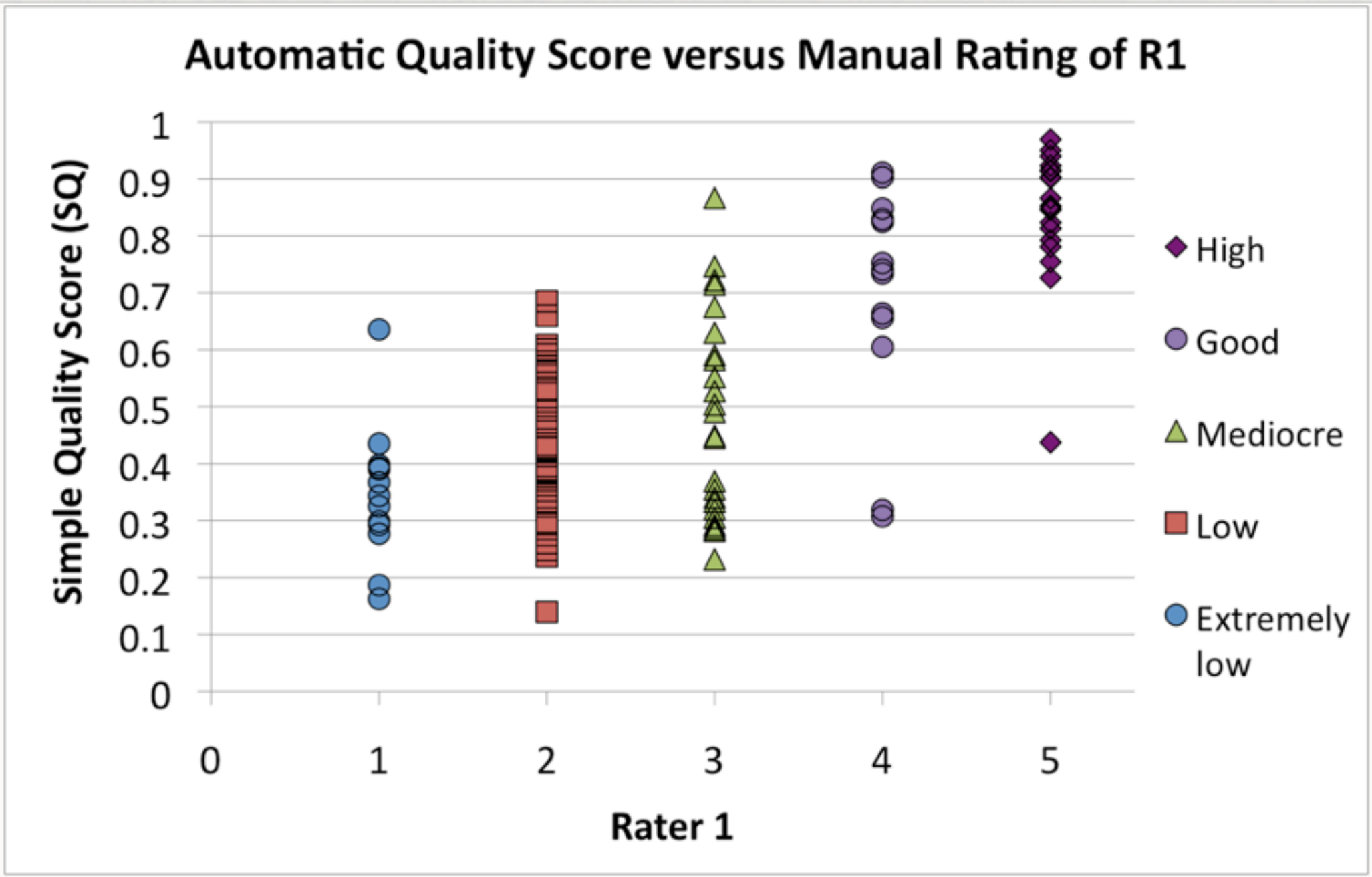
INTER-RATER AGREEMENT



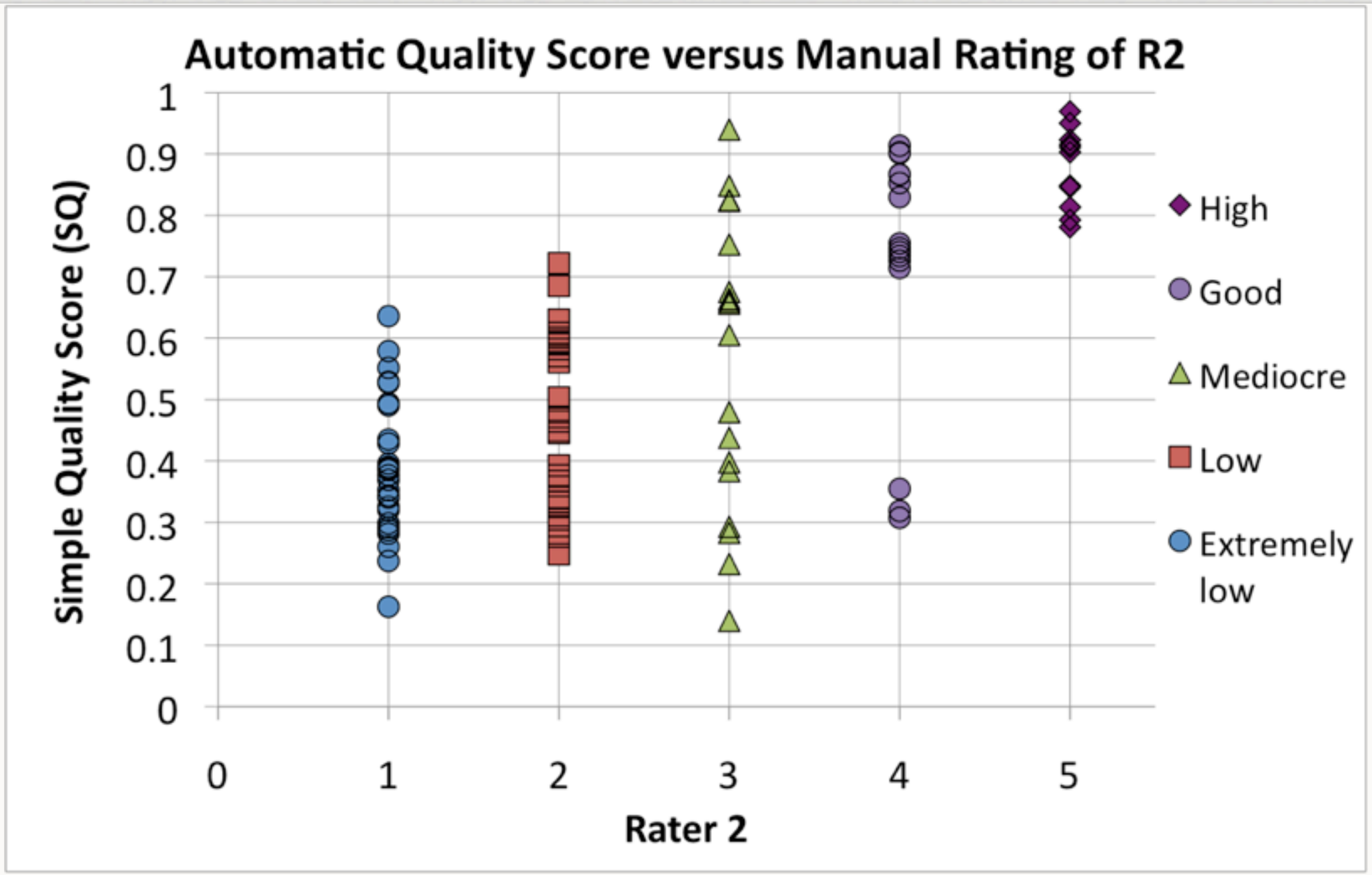
INTER-RATER AGREEMENT



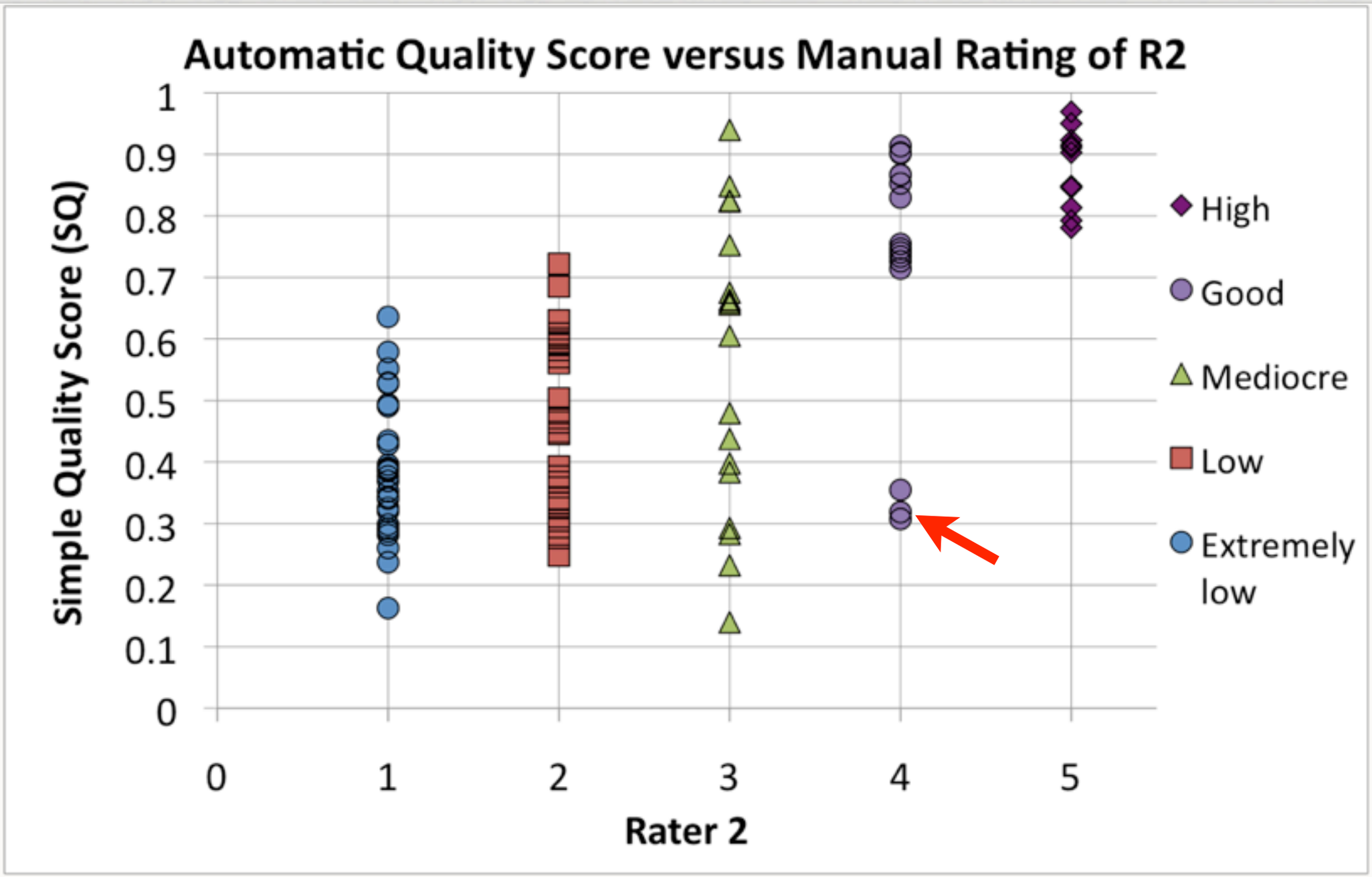
AUTOMATIC VS HUMAN



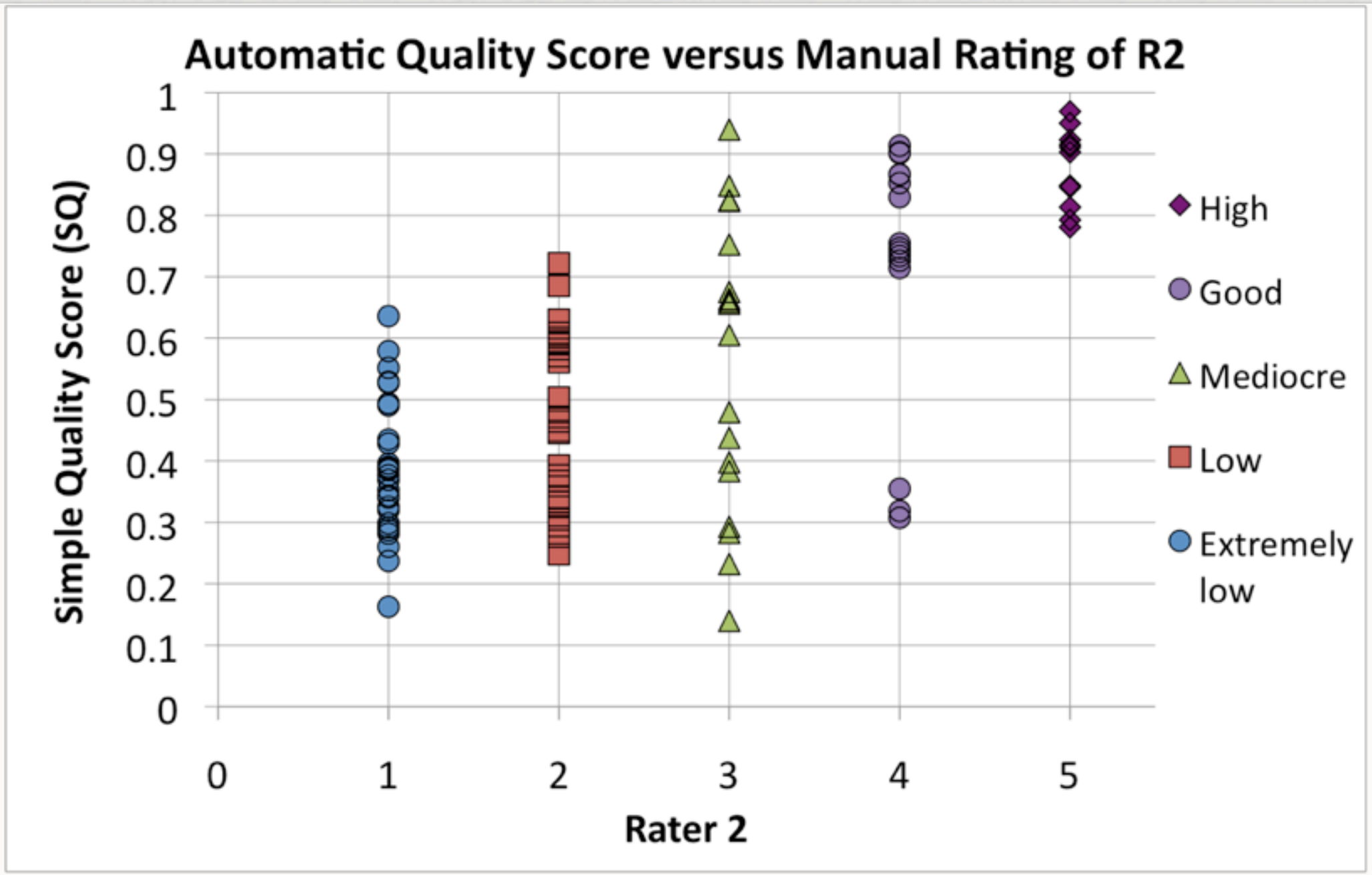
AUTOMATIC VS HUMAN



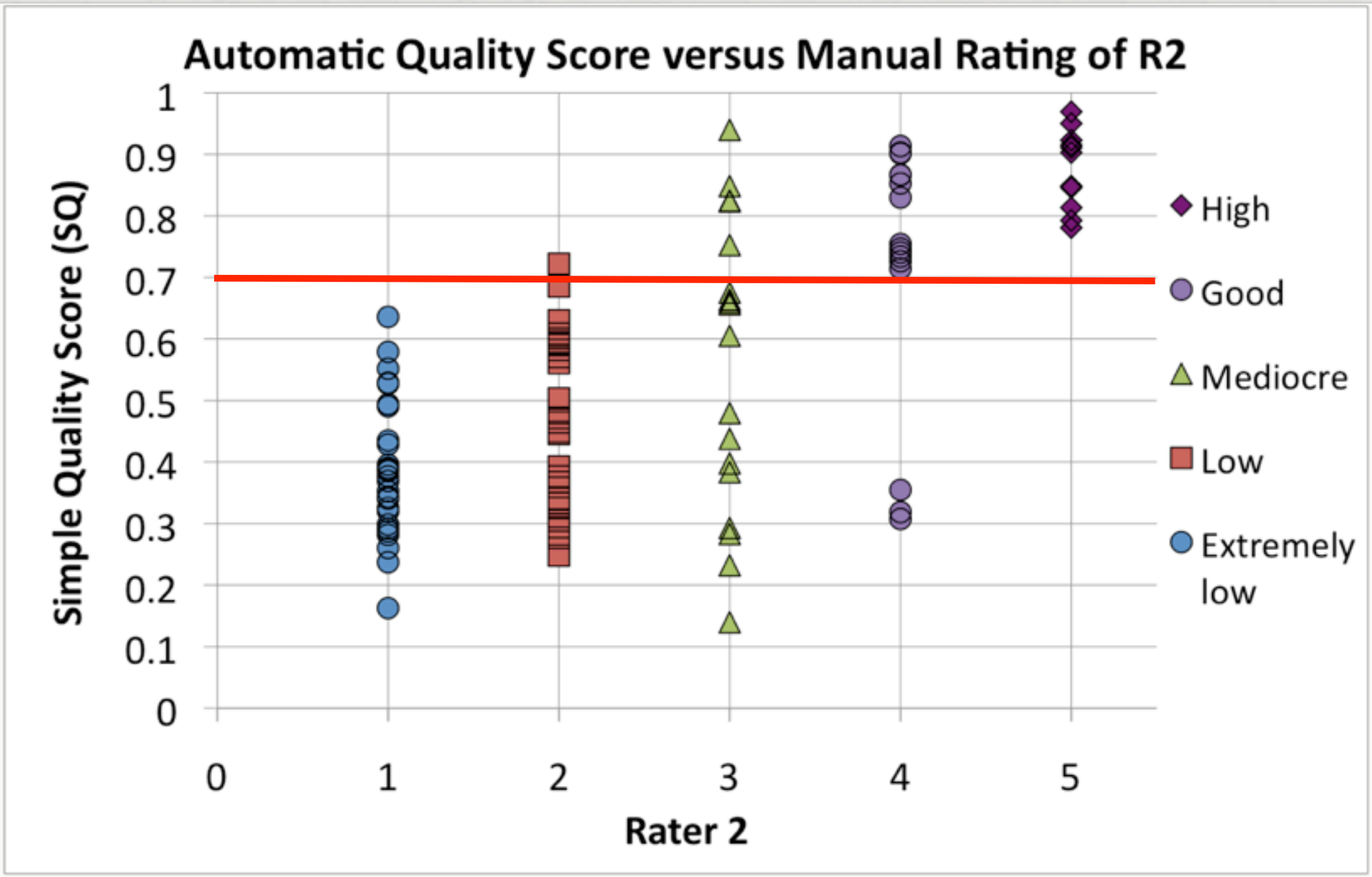
AUTOMATIC VS HUMAN



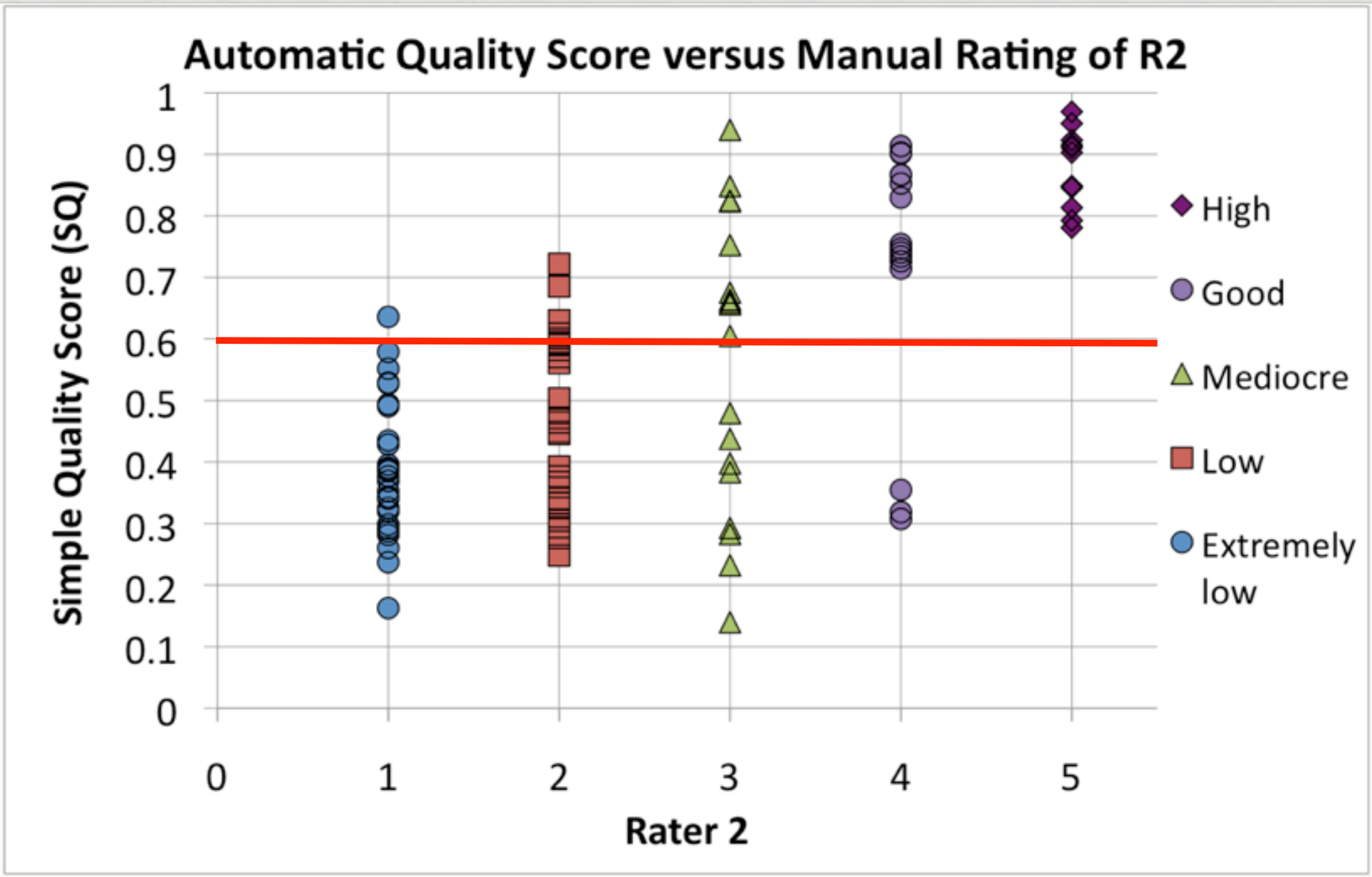
AUTOMATIC VS HUMAN



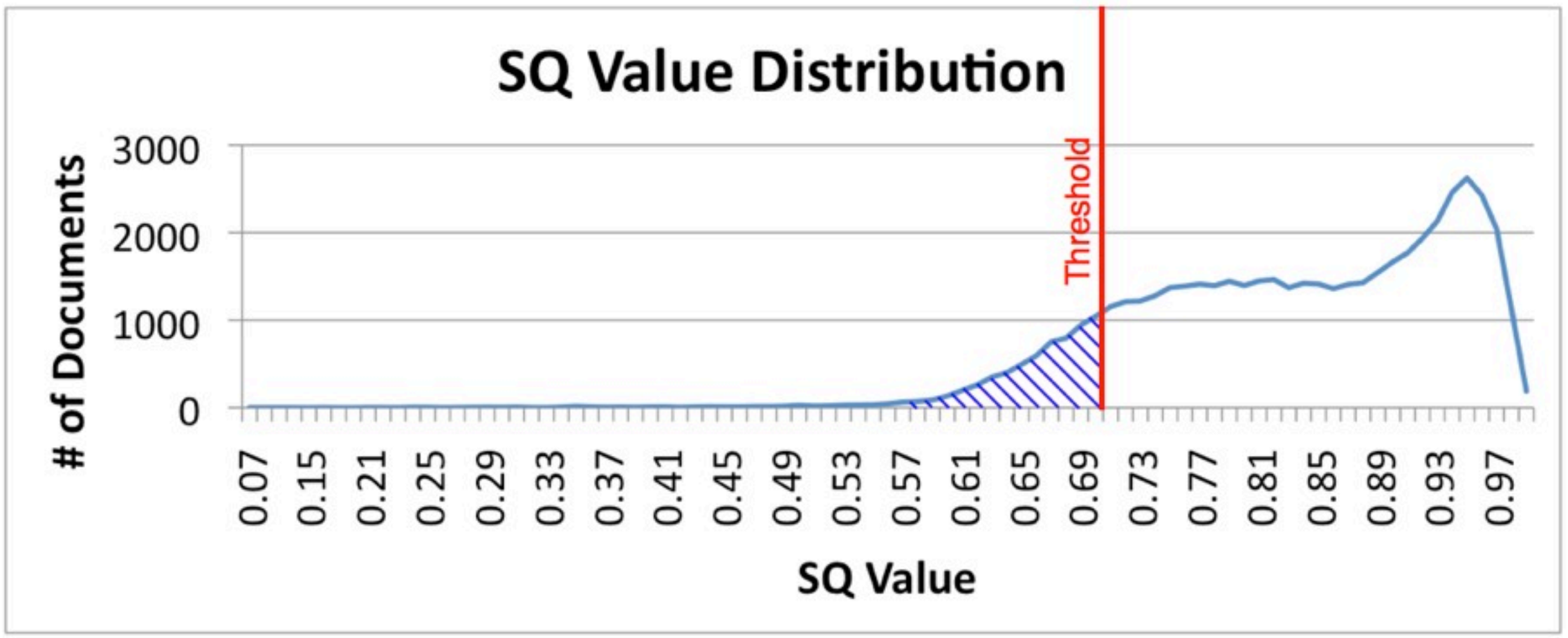
AUTOMATIC VS HUMAN



AUTOMATIC VS HUMAN



THRESHOLD?


SQ Value Distribution
# of Documents
3000
2000
1000
0
Threshold
0.07
0.15
0.21
0.25
0.29
0.33
0.37
0.41
0.45
0.49
0.53
0.57
0.61
0.65
0.69
0.73
0.77
0.81
0.85
0.89
0.93
0.97
SQ Value

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.

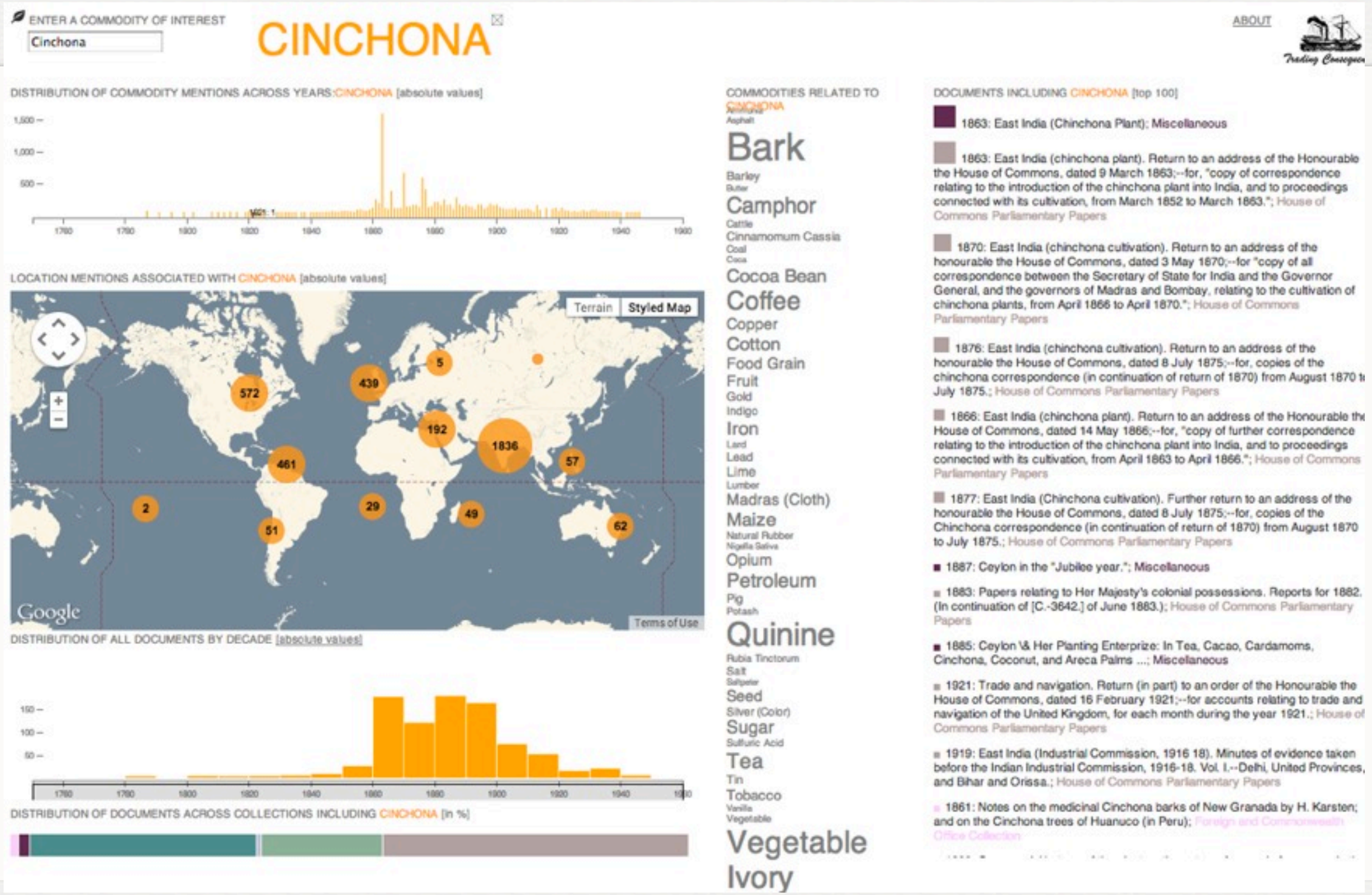
FUTURE WORK

- 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 loco-specificity and its context in literature.

THANK YOU

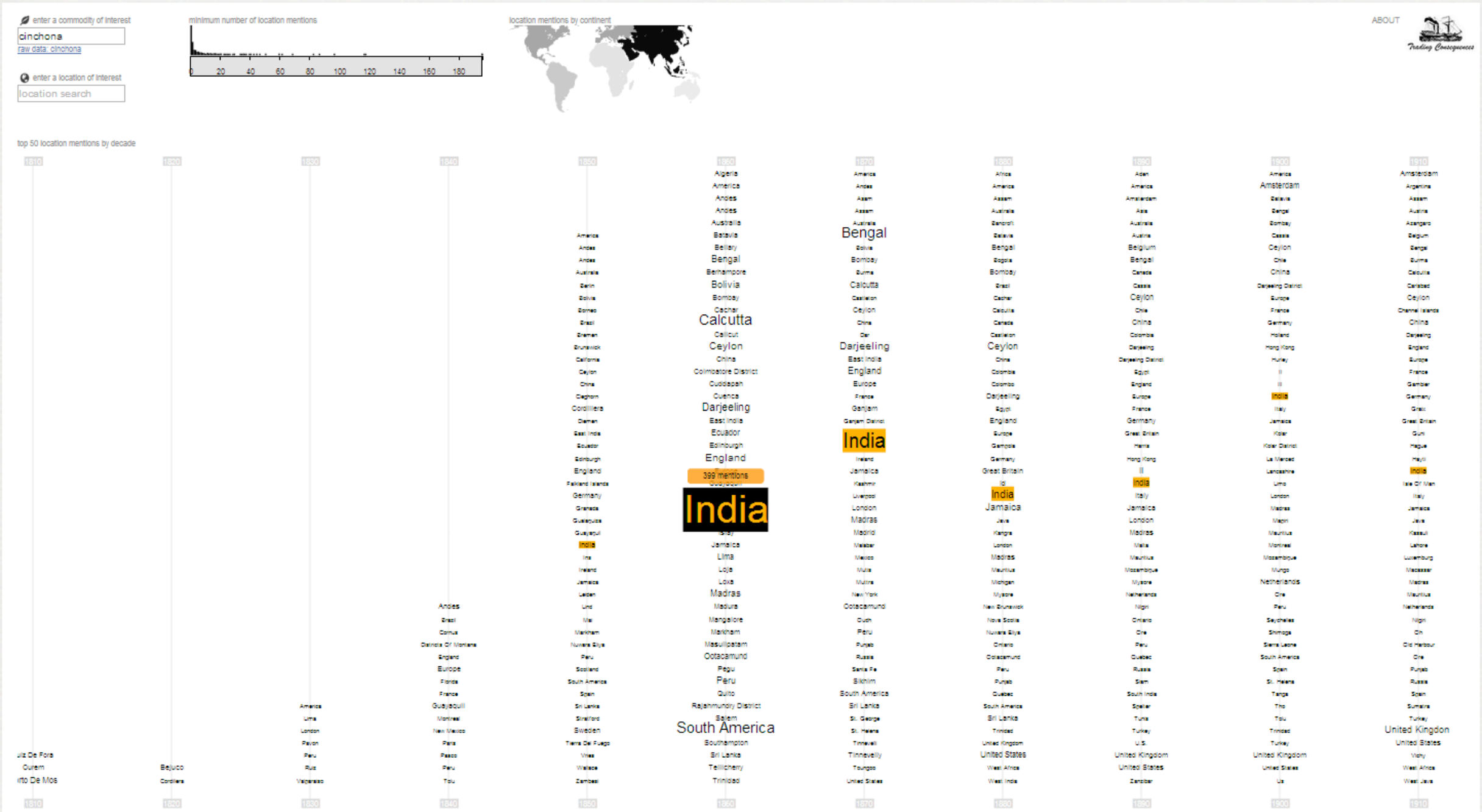
- Rating annotation guidelines and doubly rated data available on GitHub (digtrade)
- Contact: balex@inf.ed.ac.uk
- Website: http://tradingconsequences.blogs.edina.ac.uk/
- Twitter: @digtrade

BRINGING ARCHIVES ALIVE



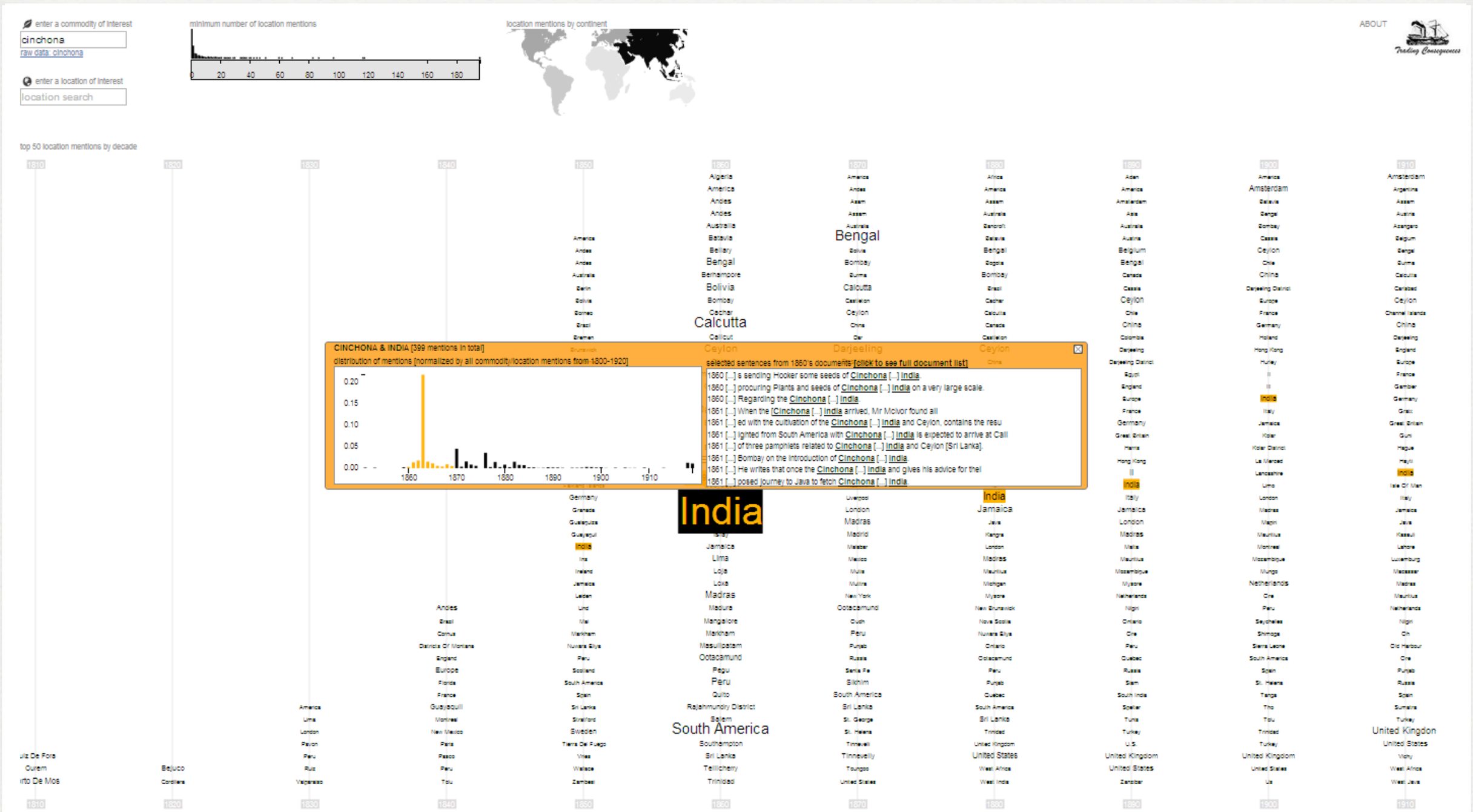
DATeCH 2014, May 20th 2014

BRINGING ARCHIVES ALIVE



DATeCH 2014, May 20th 2014

BRINGING ARCHIVES ALIVE

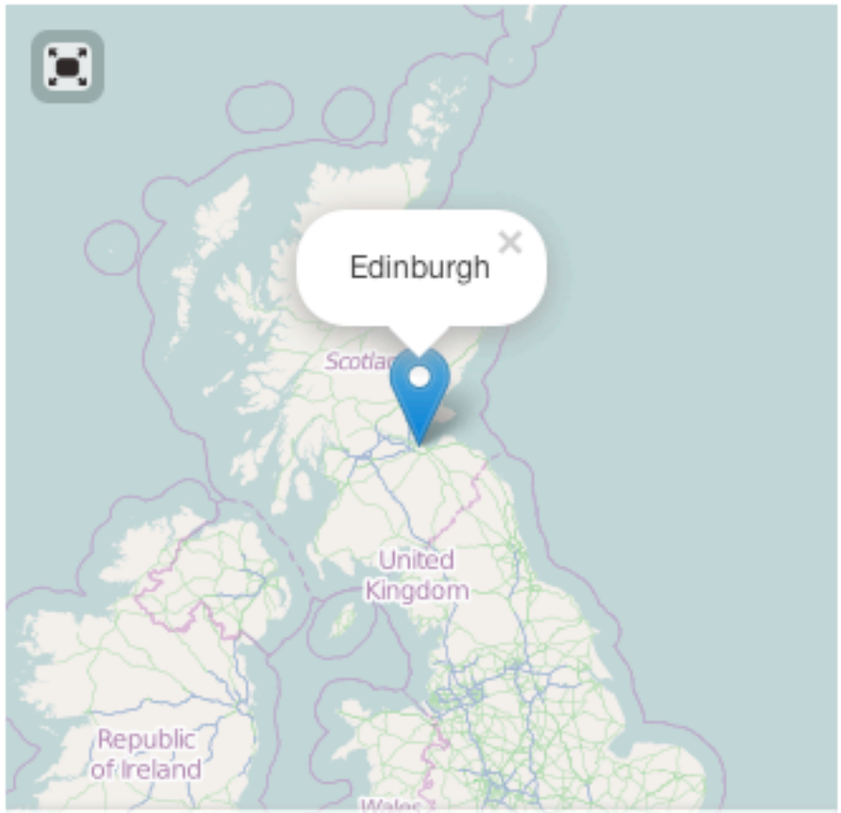


DATeCH 2014, May 20th 2014

Search by: Commodity | Location

Edinburgh

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



Filter

BY COLLECTION

All

House of Commons Parliamentary Papers (116)

BY DECADE

All

1850s (116)

BY COMMODITY

All

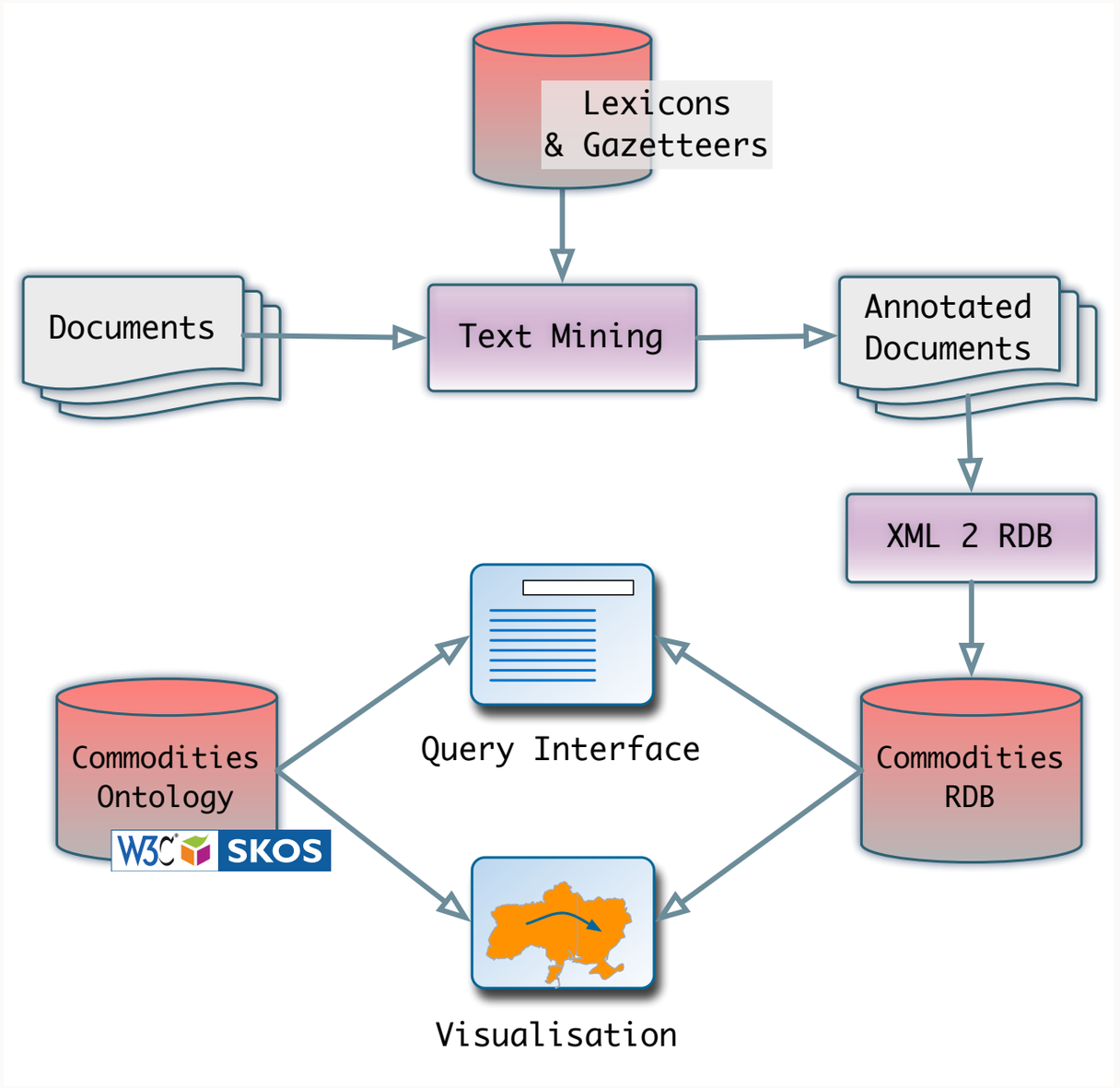
Gold (116)

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)

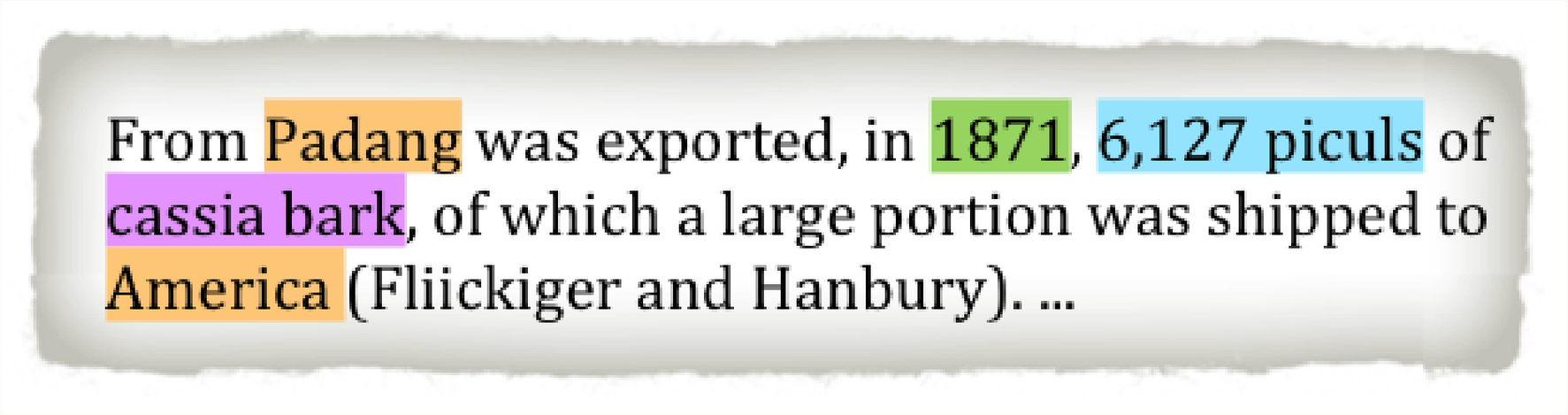
# Mentions	Document Title
90	Report from the Select Committee on the Bank Acts; together with the proceedings of the committee, minutes of evidence, appendix and index.
4	Twenty-ninth report of the Commissioners of Her Majesty's Woods, Forests and Land Revenues: in obedience to the acts of 10 George 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

SYSTEM


Lexicons
& Gazetteers
Documents
Text Mining
Annotated
Documents
XML 2 RDB
Commodities
Ontology
W3C SKOS
Query Interface
Commodities
RDB
Visualisation

MINED INFORMATION

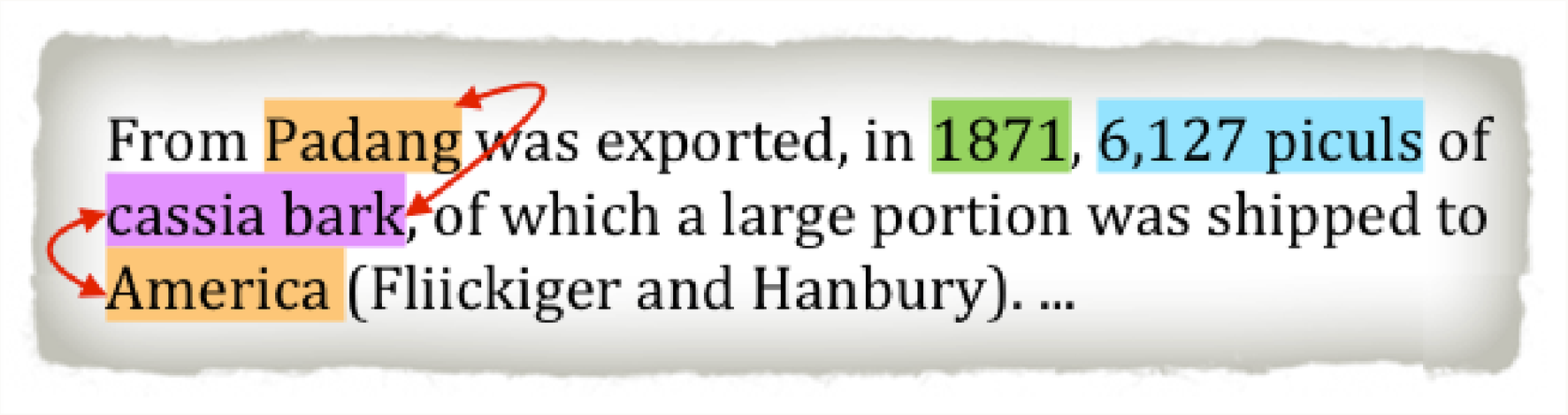
- Example sentence:



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

MINED INFORMATION

- Example sentence:



- 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	Victoria	Malaysia	seat of a first-order administrative division	10558.79 km
7	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

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

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.