Analysing Digital Historical Textual Data at Scale with Defoe Toolbox

Dr. Rosa Filgueira, EPCC,
University of Edinburgh
Email: rosa.filgueira@ed.ac.uk

Context

Working with
- Historians, Humanities and computational linguistics researchers
- Large digital collections been available for research

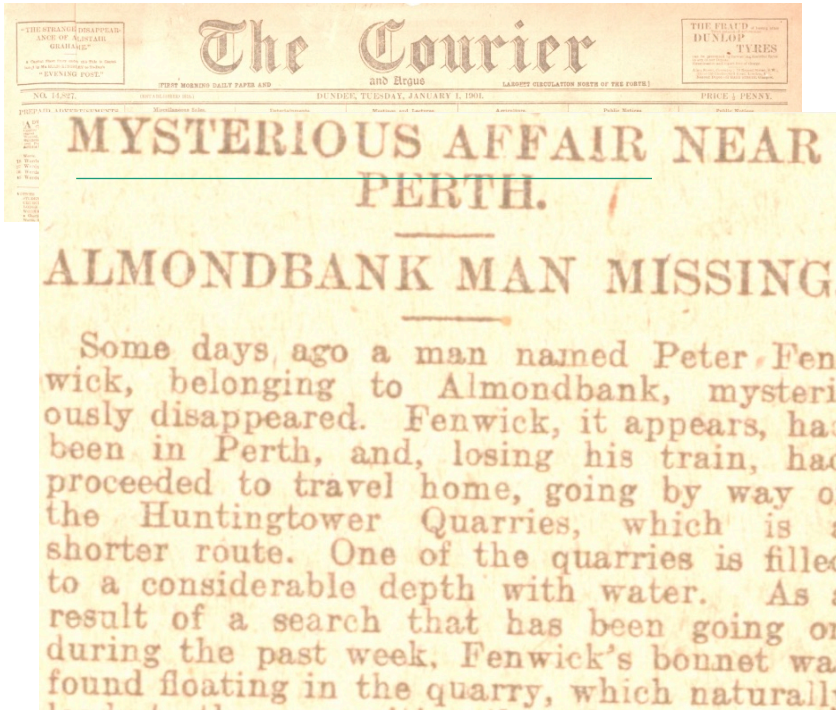
Funding - March 2018
- ATI–SE Data Engineering Programme
- Living with Machines (LwM)
- Text Data Mining (TDM)
- CDCS Text Data Mining Lab

Motivation – eScience for Historians & Humanities comunities
- Hunger for large scale text mining facilities
- Limited capacity and/or skills to use:
    - HPC/Cloud environments
    - analytic frameworks to create applications

Context

Challenges

- Several large digital collections (semi-structured data)
- Different levels of quality of data – OCR
- Data with different physical representations and schemas



The Courier and Argus

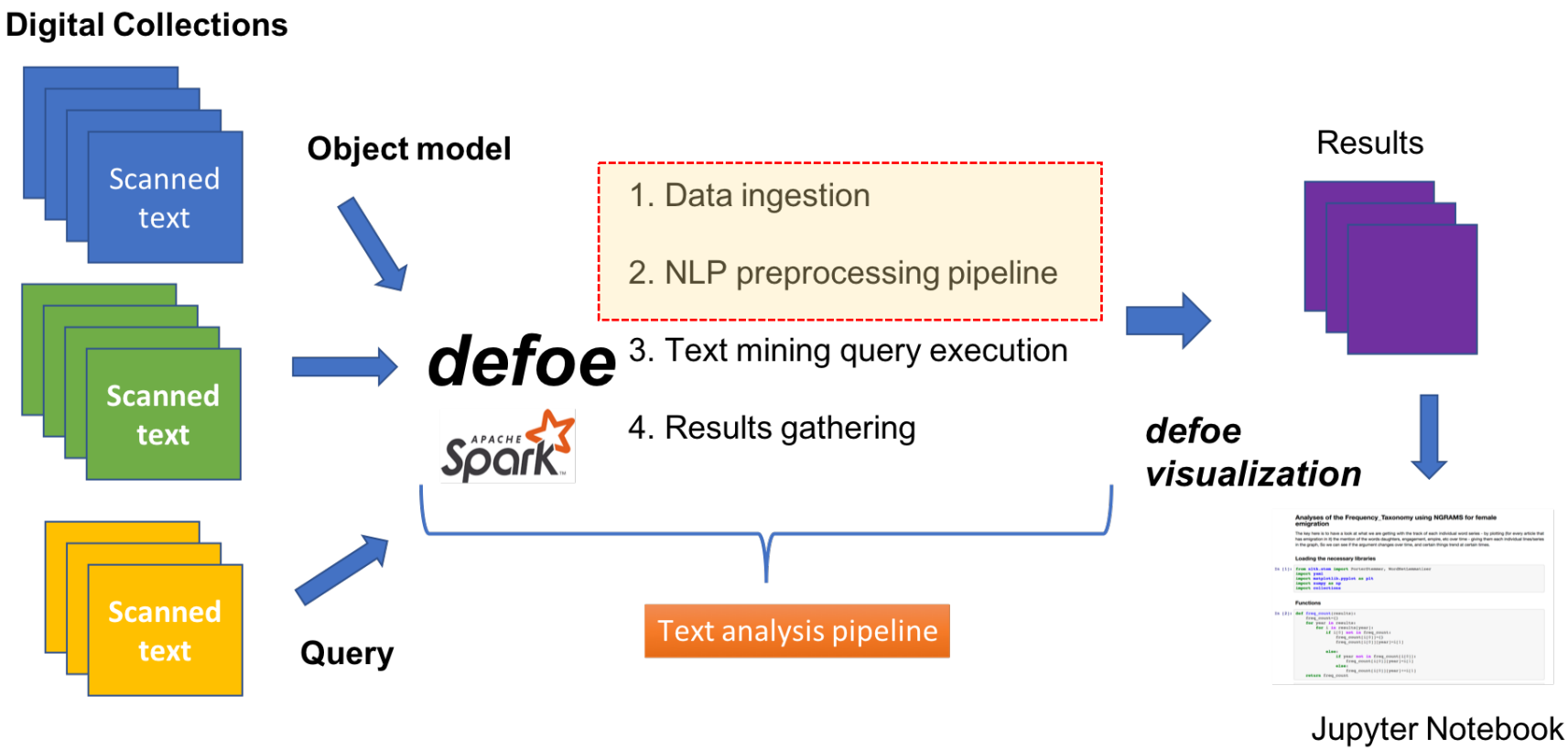
MYSTERIOUS AFFAIR NEAR PERTH.

ALMONDBANK MAN MISSING

Some days ago a man named Peter Fen wick, belonging to Almondbank, mysteri ously disappeared. Fenwick, it appears, ha been in Perth, and, losing his train, ha proceeded to travel home, going by way o the Huntingtower Quarries, which is shorter route. One of the quarries is fille to a considerable depth with water. As result of a search that has been going o during the past week, Fenwick's bonnet wa found floating in the quarry, which naturall

```
...
<text.title>
  <pg pgref="5" clipref="1"
      pos="4069,3036,4949,3154"/>
  <p>
    <wd pos="4069,3036,4949,3154">MYSTERIOUS AFFAIR
NEAR PERTH.</wd>
  </p>
</text.title>
<text.cr>
  <pg pgref="5" clipref="1"
      pos="4039,3191,4987,4235"/>
  <p>
    <wd pos="4041,3192,4496,3241">ALMONDBANK</wd>
    <wd pos="4523,3200,4663,3246">MAN</wd>
    <wd pos="4696,3198,4976,3250">MISSING.</wd>
    <wd pos="4085,3290,4189,3323">Some</wd>
    <wd pos="4214,3290,4312,3329">days,</wd>
    ...
```

defoe: new eScience toolbox for historical research

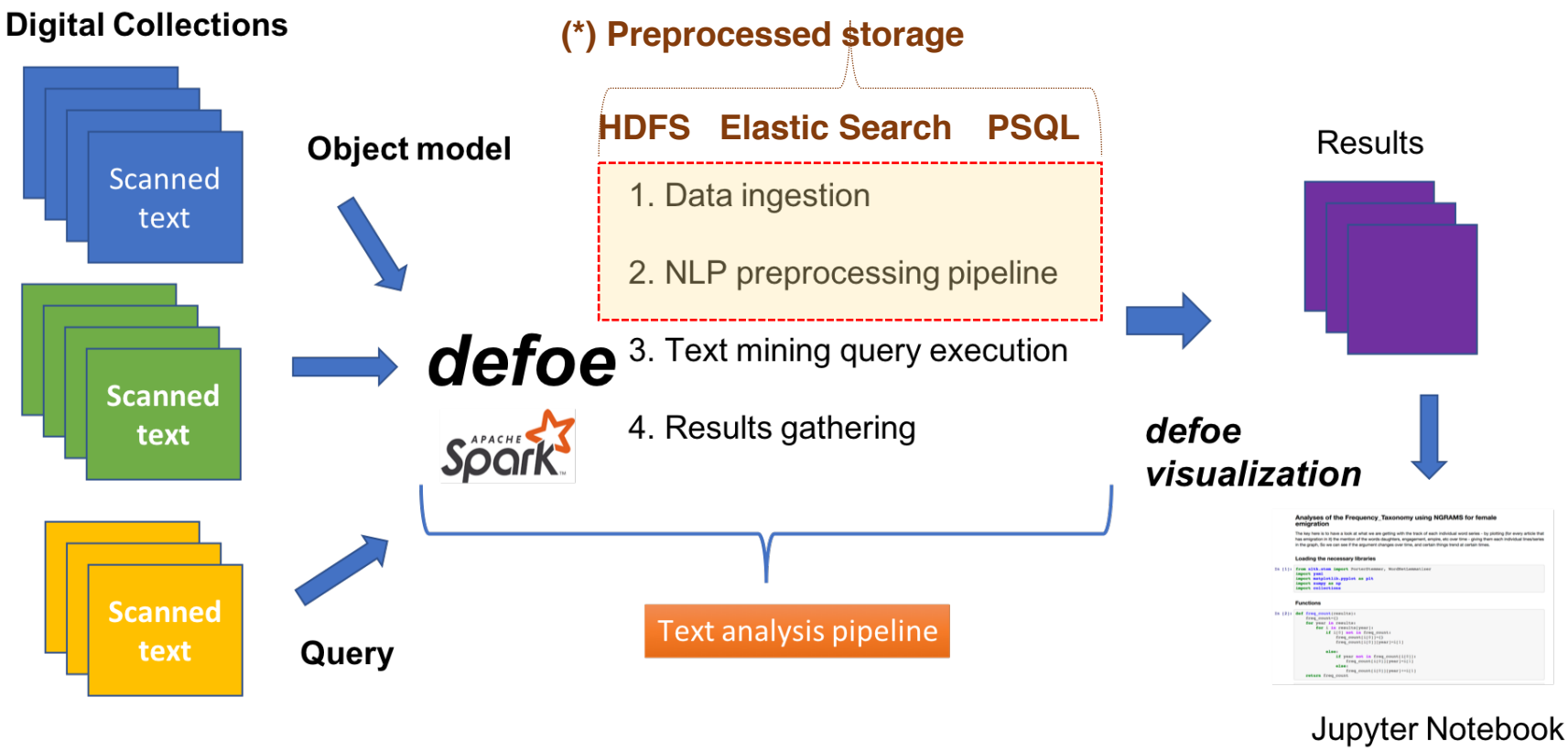


scalable portable text mining queries large dataset

https://github.com/alan-turing-institute/defoe

https://github.com/alan-turing-institute/defoe_visualization

defoe: new eScience toolbox for historical research

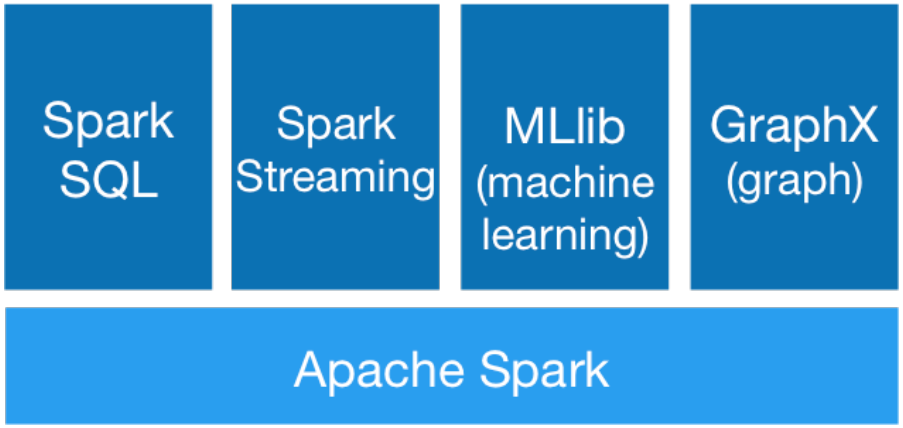


scalable portable text mining queries large dataset

https://github.com/alan-turing-institute/defoe

https://github.com/alan-turing-institute/defoe_visualization

Apache Spark



https://spark.apache.org/

- Analytics engine for large-scale data processing
- High performance for batch and streaming data
- APIs

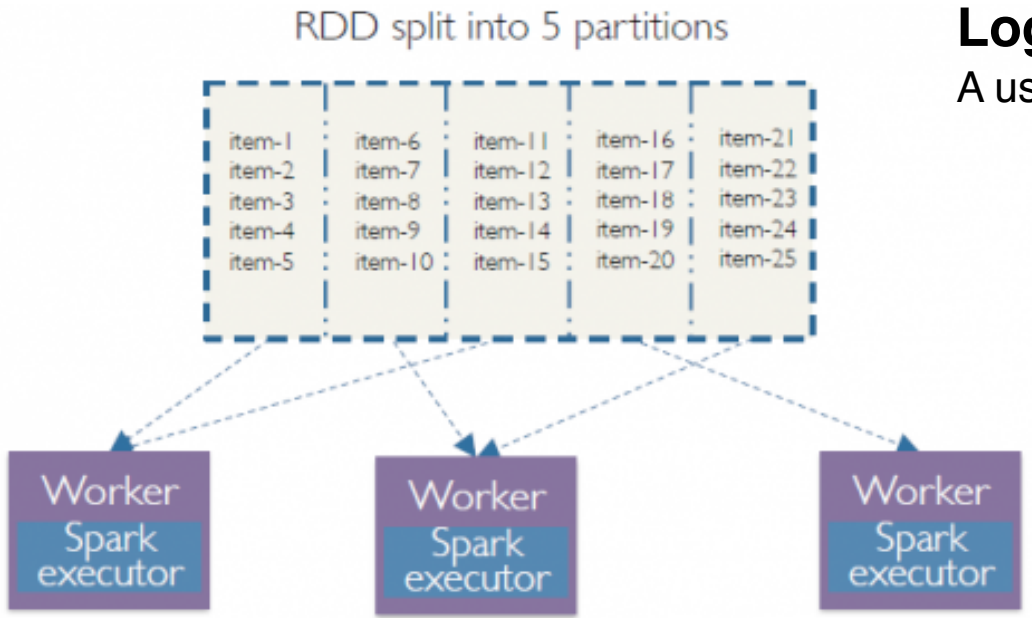
  Java, Scala, **Python** and R

https://github.com/EPCCed/prace-spark-for-data-scientists

Apache Spark

Resilient Distributed Datasets (RDD)

- Represent data or transformations on data
- It is distributed collection of items – partitions
- Read-only → they are immutable
- Enables operations to be performed in parallel



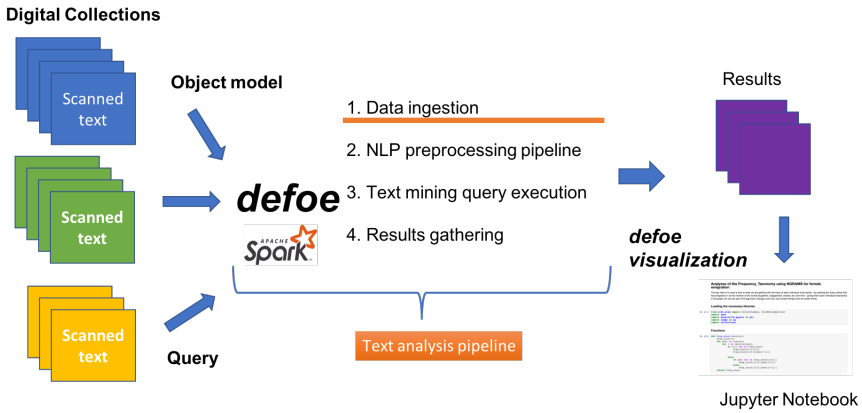
Logical – view

A user sees an object with 25 items

Physically – view

25 items distributed across different CPUS/Cores

Data Ingestion



Support for three physical representations:

1) one XML document per issue
2) one XML document with search results including several articles
3) one XML METS document and ALTO XML per page

Object models -- **loading/parsing** data into RDD:

- **PAPER**
- NZPP
- ALTO
- FMP
- **NLS**

PAPER object model (British Library Newspapers)

0000164- The Courier and Argus
0000187- The Bath Chronicle
0000195- Archer Bath Chronicle
0000321- The Nottingham Evening Post
0000452- Edinburgh Evening News

RDD

Class Issue → Representation of an issue (XML document)
Each XML holds articles that belong to the same issue

issue

attributes

- filename
- issue_tree
- issue_id
- date
- page_count
- day_of_the_week
- **article list**

Class Article → Representation of an article

article

attributes

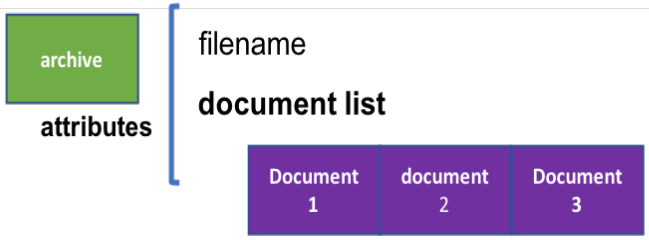
- article_tree
- filename
- quality
- title
- preamble
- content
- article_id
- page_ids
- words = content + title +preamble

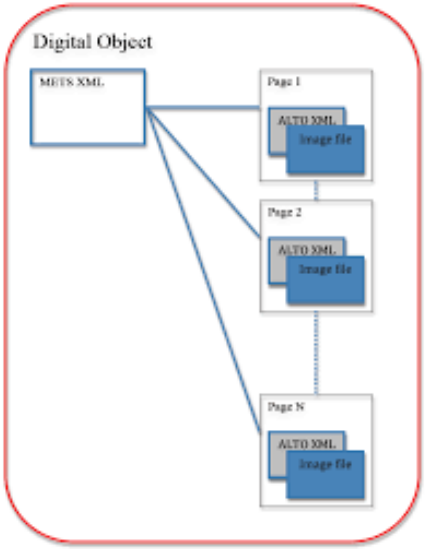
ALTO/NLS object models (British Library books and NLS Digital Collection)

RDD

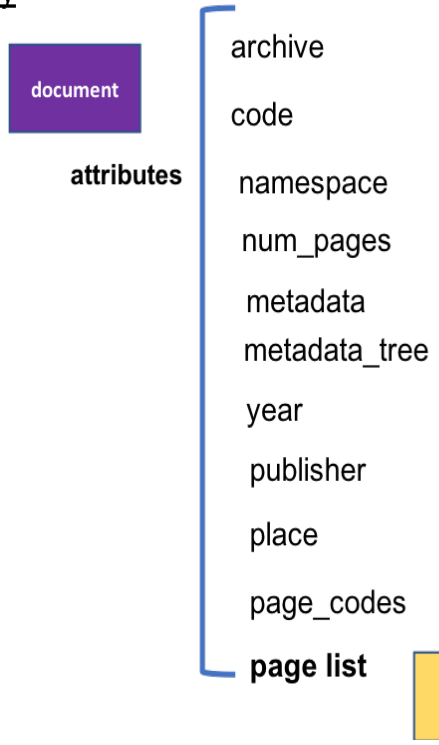
1510_1699/000001143_0_1-20pgs__560409_dat.zip
1510_1699/000000874_0_1-22pgs__570785_dat.zip
1510_1699/000051983_0_1-92pgs__568584_dat.zip
1510_1699/000987728_0_1-92pgs__567840_dat.zip

Class Archive → Representation of a zipped archive

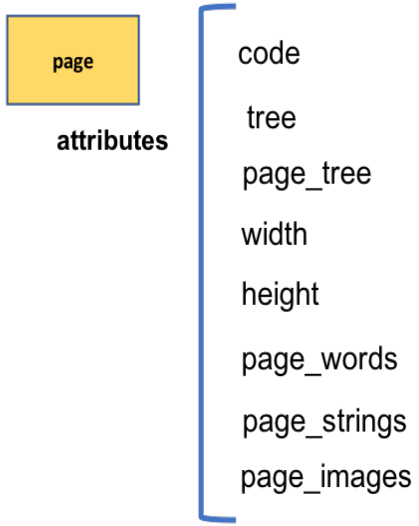




Class Document → Representation of a metadata document (XML in METS/MODS) and an ALTO directory



Class Page → Representation of a page (XML in ALTO).

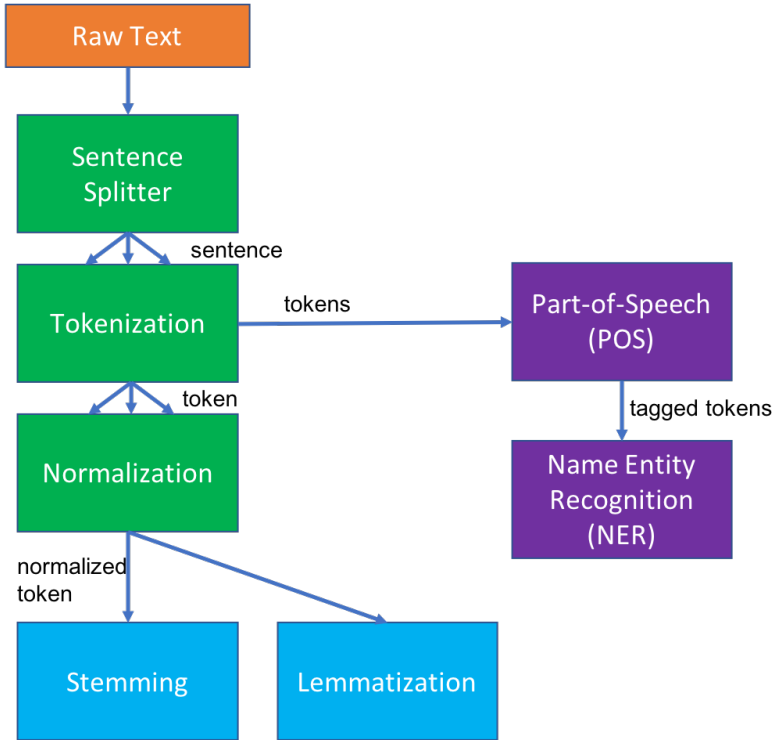


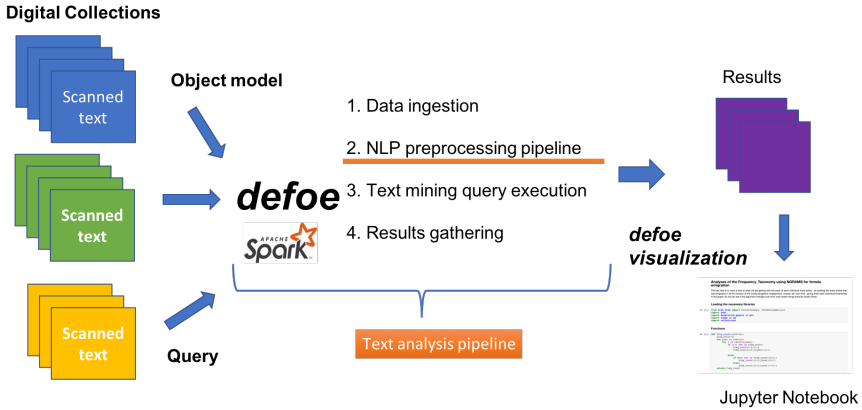
Digital Collections

Dataset	Period	Structure	XML Schema	Space	Model
British Library Books (BLB)	1510-1899	ZIP per book - XML metadata - XML per page	METS and ALTO schemas	~220GB	**ALTO**
British Library Newspapers (BLN)	1714-1950	XML per issue	GALEN Schema	~1TB	**PAPERS**
Times Digital Archive (TDA)	1785-2009	XML per issue	GALEN Schema	~324GB	**PAPERS**
Papers Past New Zealand and Pacific newspapers (NZPP)	1839-1863	XML per 22 articles	XML from a search via an API	~4GB	**NZPP**
Gazetteers of Scotland, 1803-1901 (*)	1803-1901	- XML metadata - XML per page	METS and ALTO schema	4.7 GB (includ. Fig)	**NLS**
Encyclopaedia Britannica (**)	1768-1860	- XML metadata - XML per page	METS and ALTO schemas	46GB (includ. figs)	**NLS**

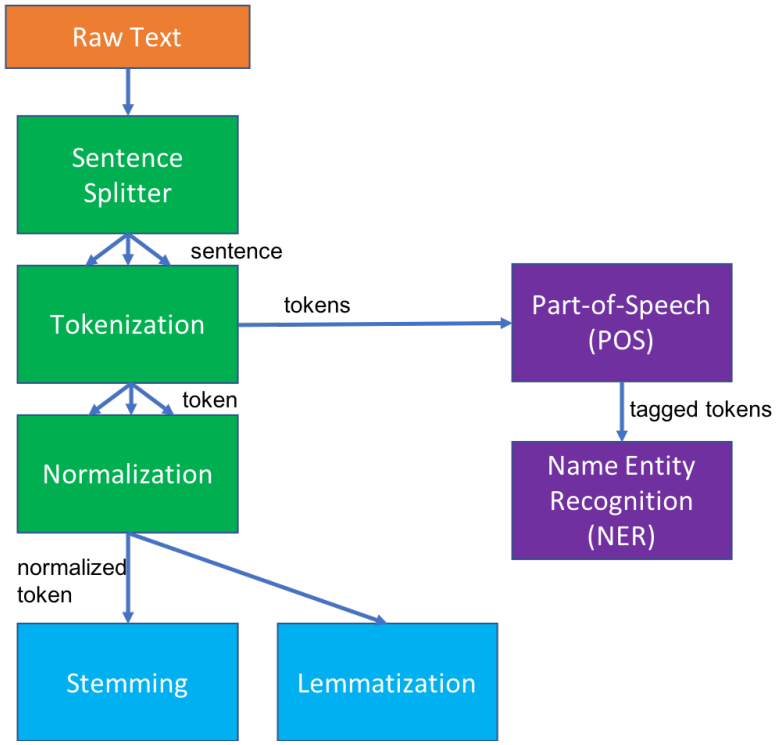
(*) https://data.nls.uk/data/digitised-collections/gazetteers-of-scotland/
(**)https://data.nls.uk/data/digitised-collections/encyclopaedia-britannica/

NLP Preprocessing





NLP Preprocessing



1. Divide the text in sentences

2. Each word of a sentence is tokenised

3. Each token is normalised

4.a) Stemming: reduces words to their roots

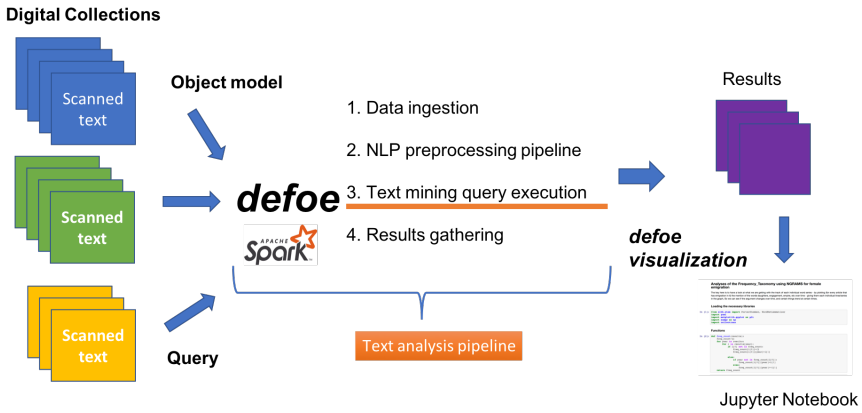
4.b) Lemmatisation: reduces words to a common base

5. Tag each token
   (e.g. nouns, verbs, adjectives, etc.)

6. Classify tagged tokens
   (e.g. names of persons, locations, etc)

Spacy and NLTK

Text Mining Queries



NLS Model

- colocates_by_year.py
- geoparser_pages.py
- georesolution_pages.py
- inventory_per_year.py
- keysearch_by_year.py
- keysentence_by_year_paper.py
- keyword_by_word.py
- keyword_by_year.py
- keyword_concordance_by_word.py
- keyword_concordance_by_year.py
- keyword_metadata_by_word.py
- normalize.py
- ocr_quality_by_year.py
- ocr_quality_multi_level_by_year.py
- preprocessing_sentences.py
- total_documents.py
- total_pages.py
- total_words.py
- write_pages_df_es.py
- write_pages_df_hdfs.py
- write_pages_df_psql.py

ALTO Model

- colocates_by_year.py
- keyword_by_word.py
- keyword_by_year.py
- keyword_concordance_by_word.py
- keyword_concordance_by_year.py
- normalize.py
- ocr_quality_by_year.py
- total_documents.py
- total_pages.py
- total_words.py

PAPERS Model

- colocates_by_year.py
- keysentence_by_year.py
- keyword_by_year.py
- keyword_concordance_by_date.py
- keywords_by_year.py
- lda_topics.py
- normalize.py
- ocr_quality_by_year.py
- target_and_keywords_by_year.py
- target_and_keywords_count_by_ye...
- target_concordance_collocation_b...
- total_articles.py
- total_issues.py
- total_words.py
- unique_words.py

NZPP Model

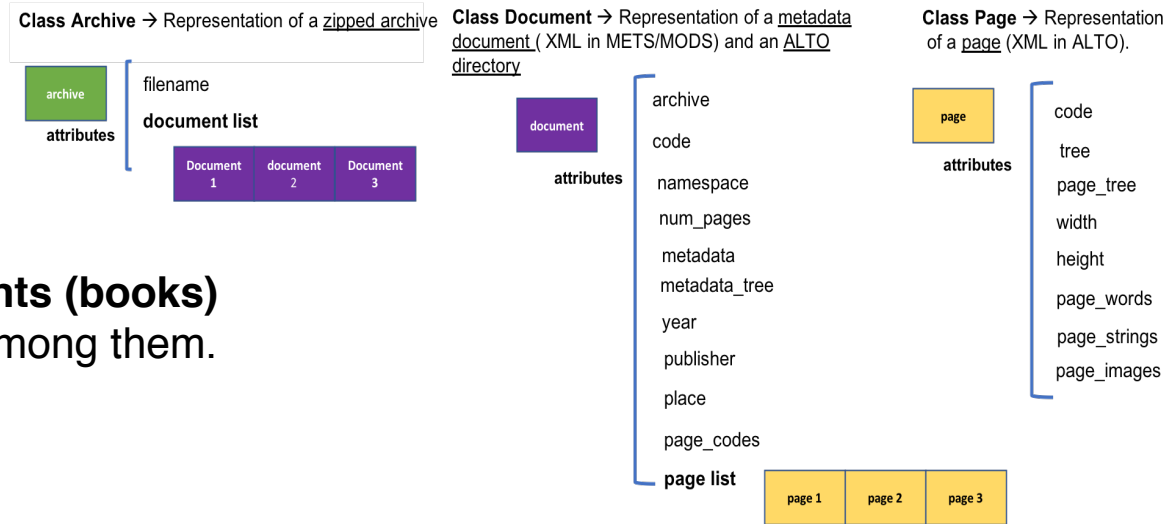
- keyword_by_year.py
- keyword_concordance_by_date.py
- normalize.py
- total_articles.py
- total_words.py

FMP Model

- keyword_metadata_by_word.py
- keyword_segmentation.py
- normalize.py
- target_segmentation.py
- total_articles.py
- total_documents.py

Text Mining Queries

ALTO object model



total_words query:

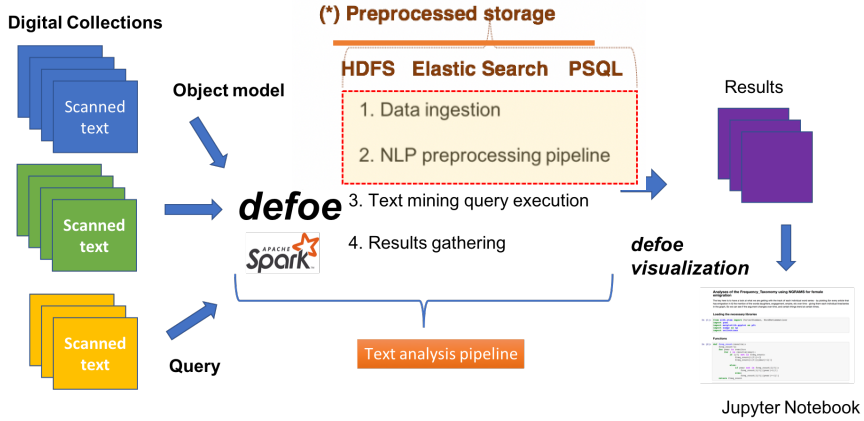
- Iterates through archives
- Count **total number of documents (books)** and **total number of words** among them.

Sample results

Query over British Library Books

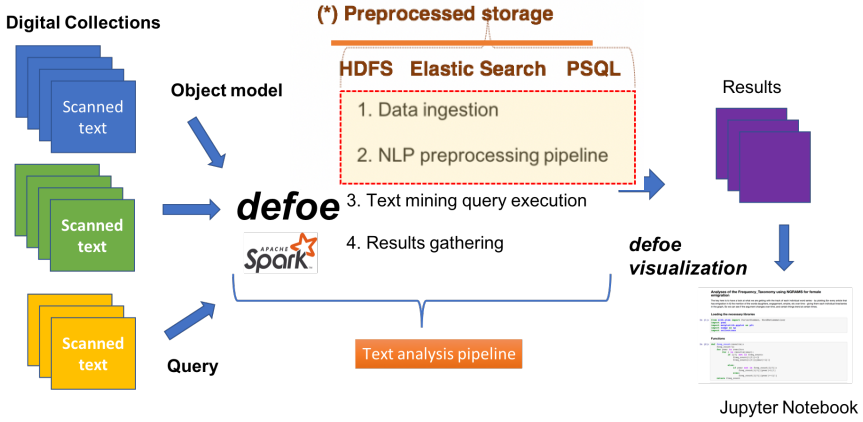
```
{num_documents: 63701, num_words: 6866559285}
```

Preprocessed Data



1. Transform collections to a common format after being preprocessed
2. Run the same queries across collections
3. Capture of metadata + Raw Text + Preprocessed text

Preprocessed Data



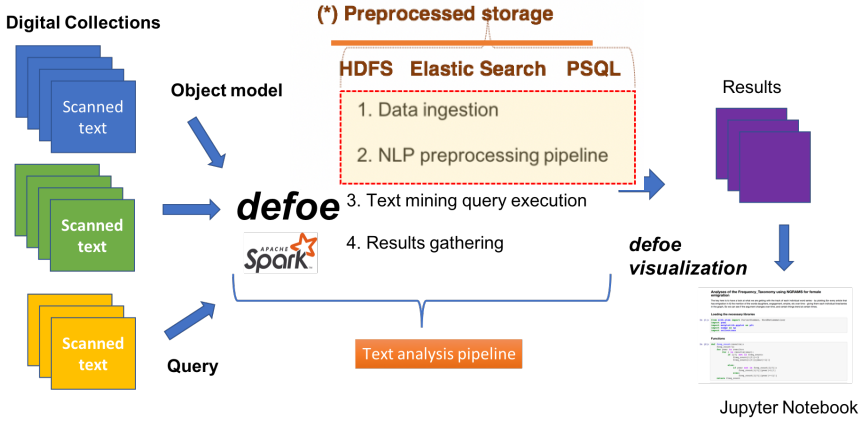
Title: Title of the collection - e.g. *Encyclopaedia Britannica*
Edition: Edition of the book/issue - e.g. *Seventh edition, Volume 13, LAB-Magnetism*
Year: Year of publication/edition - e.g. *1842*
Place: Place - e.g. *Edinburgh*

archive_filename: Directory Path of the book/newspaper - *e.g. /lustre/home/ …/…/193108323*
source_text_filename: Directory Path of the page/issue - e.g. *alto/193201316.34.xml*

type_archive: Type of archive *book / newspaper*
text_unit: Unit that represent each ALTO XML - e.g *Page/Article*
num_text_unit: Number of pages of a book / Number of articles of an issue - *e.g. 810*
text_unit_id: Id of the page - e.g. *Page3*
model: defoe model used for ingesting this dataset – e.g *nls*

source_text_raw: Text of a page/article
source_text_clean: Cleaned text of page/article – long-S and hyphened word fixes
source_text_norm: Normalised cleaned text
source_text_lemmatized: Lemmatized normalized cleaned text
source_text_stemm: stemmed normalized cleaned
num_words: Number of words of a page/article

Preprocessed Data

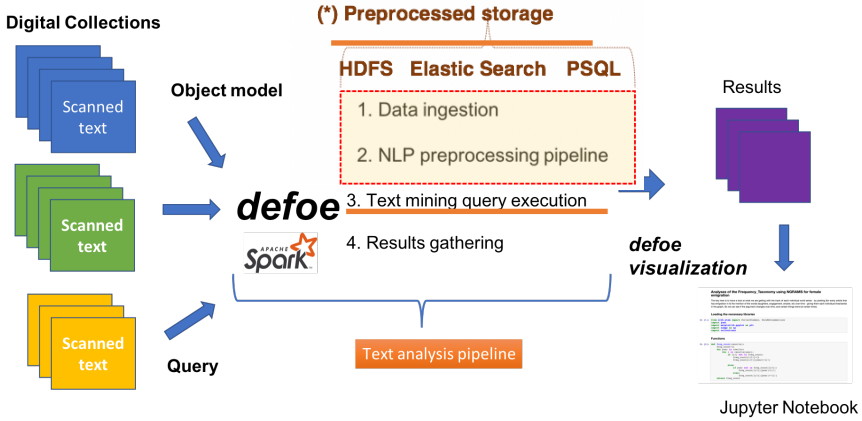


Example - Scottish Gazetteers

title	edition	year	place	archive_filer	source_text_	text_unit	text_unit_id	num_text_u	type_archive	model	source_text_raw	source_text_clean
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973434	page	Page1	606	book	nls		
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973434	page	Page2	606	book	nls	iA<* R&o: ;,¬´.'-- ',Ä¢,Ä¢- ;\<,Ä¢\""	i A<* R&o: ;,¬´.'-',Äc,Ä¢;\<,Ä¢\""
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973434	page	Page3	606	book	nls	^L o,	^L o,
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973434	page	Page4	606	book	nls		
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973434	page	Page5	606	book	nls	Digitized by the Internet Archive in 2012 http://archive.org/details,	Digitized by the Internet Archive in 2012 http://
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page6	606	book	nls		
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page7	606	book	nls	Jac:7T. /. 2 1 tttttfUvCGZl cu&cvtcJ	Jac:7T. /. 2 1 tttttf Uv CG Zl cu&cvtc J
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page8	606	book	nls		
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page9	606	book	nls	THEl GAZETTEER, OF SCOTLAND; CONTAINING A PARTICULAR ANI	TH El GAZETTEER, OF SCOTLAND; CONTAINING
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page10	606	book	nls		
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page11	606	book	nls	INTRODUCTION. GENERAL DESCRIPTION. SGOTLAND, or that part	INTRODUCTION. GENERAL DESCRIPTION. SGO
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page12	606	book	nls	iv INTRODUCTION. great opening is termed the Moray Frith, and e	iv INTRODUCTION. great opening is termed the
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page13	606	book	nls	INTRODUCTION. v acres of cultivated, and 14,218,224 acres of unc	\" fays a late author
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973435	page	Page14	606	book	nls	W iNTRODUCTION. Lammermuir hills, in Berwickfhire ; the great r	4030 Benlawers
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page15	606	book	nls	INTRODUCTION. vii rising in the mountainous diftrict of Lochaber,	INTRODUCTION. vii rising in the mountainous d
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page16	606	book	nls	tin INTRODUCTION. the Atlantic ocean. Clouds of the fea are fraug	in the Highlands of Scotland
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page17	606	book	nls	INTRODUCTION. Is Highland 'Society of Scotland, who offer premi	\" fays this almoft enthufiaftic writer
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page18	606	book	nls	y- introduction: End of England, until it were heard, and underftooc	rye
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page19	606	book	nls	INTRODUCTION. 4 fiifnes, containing coins of Scottifh gold, were p	INTRODUCTION. 4 fiifnes, containing coins of S
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page20	606	book	nls	xii INTRODUCTION. fome of them of great beauty and value. Chalc	xii INTRODUCTION. some of them of great bea
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page21	606	book	nls	INTRODUCTION. xiii land, we may mention the colley, or true Shep	teem with abundance of trout
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973436	page	Page22	606	book	nls	%ly INTRODUCTION* RELIGION. It is generally believed, upon the a	%ly INTRODUCTION* RELIGION. It is generally
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page23	606	book	nls	introduction: XT Pre/byterles, and Kirk Sejfions. ift, The General AJ	introduction: XT Pre/byterles, and Kirk Sejfions.
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page24	606	book	nls	*vi INTRODUCTION. Befides the eftablifned churches, there are a v	nor aboVe two hund- red
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page25	606	book	nls	INTRODUCTION\. jcvii diftrict	by the Society for propagating the Chriftian rel
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page26	606	book	nls	xvili INTRODUCTION\. fophical works of Boethius is purely claffica	and thedatinity of Buchaniiari fe the moft claf
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page27	606	book	nls	INTRODUCTION. xi:c (ion from his Majefty of 200l. per annum. Eve	INTRODUCTION. xi:c (ion from his Majesty of 2
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page28	606	book	nls	xx INTRODUCTION. of their being a defpifed people. The great Mr.	xx INTRODUCTION. of their being a despised pe
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page29	606	book	nls	INTRODUCTION. xxi linen, and engages many thoufands of hands.	844!\") \" In -< I7^' >the quantity diftilled was-
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973437	page	Page30	606	book	nls	xxii INTRODUCTION. on thofe who raife the greateft quantities of	xxii INTRODUCTION. on those who raise the gr
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973438	page	Page31	606	book	nls	INTRODUCTION. *xi* CONSTITUTION. The ancient conftitutiori and	I promife thefe three things to the Chriftian pe
gazetteer of Scotland	1803	1803	Dundee	/lustre/hom	alto/973438	page	Page34	606	book	nls	xx3i i' n T R D U £ Ti N model of the French parliament, to ftymol	for minifters itinerds

Preprocessed Data

+ Text Mining queries



HDFS / ES/ PSQL

- georesolution_pages.py
- keysearch_by_year.py
- keysentence_concordance_by_yea...
- normalize.py
- window_concordance_by_date.py
- keysearch_by_year.py

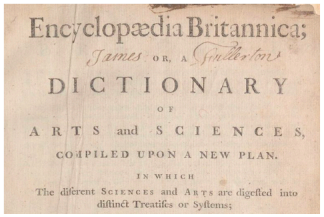
National Library Scottish – Uses Cases

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

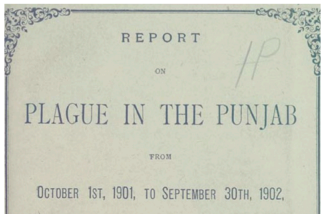
HOME ABOUT DATA TOOLS PROJECTS CONTACT

Digitised collections

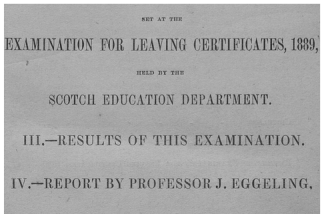
Download the ALTO, METS, image and plain text files for our digitised collections.



Encyclopaedia Britannica, 1768-1860

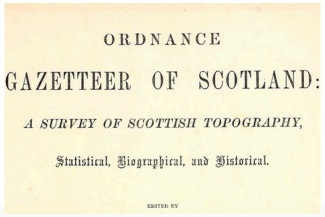


A Medical History of British India



Scottish School Exam Papers, 1888-1963

The Spiritualist



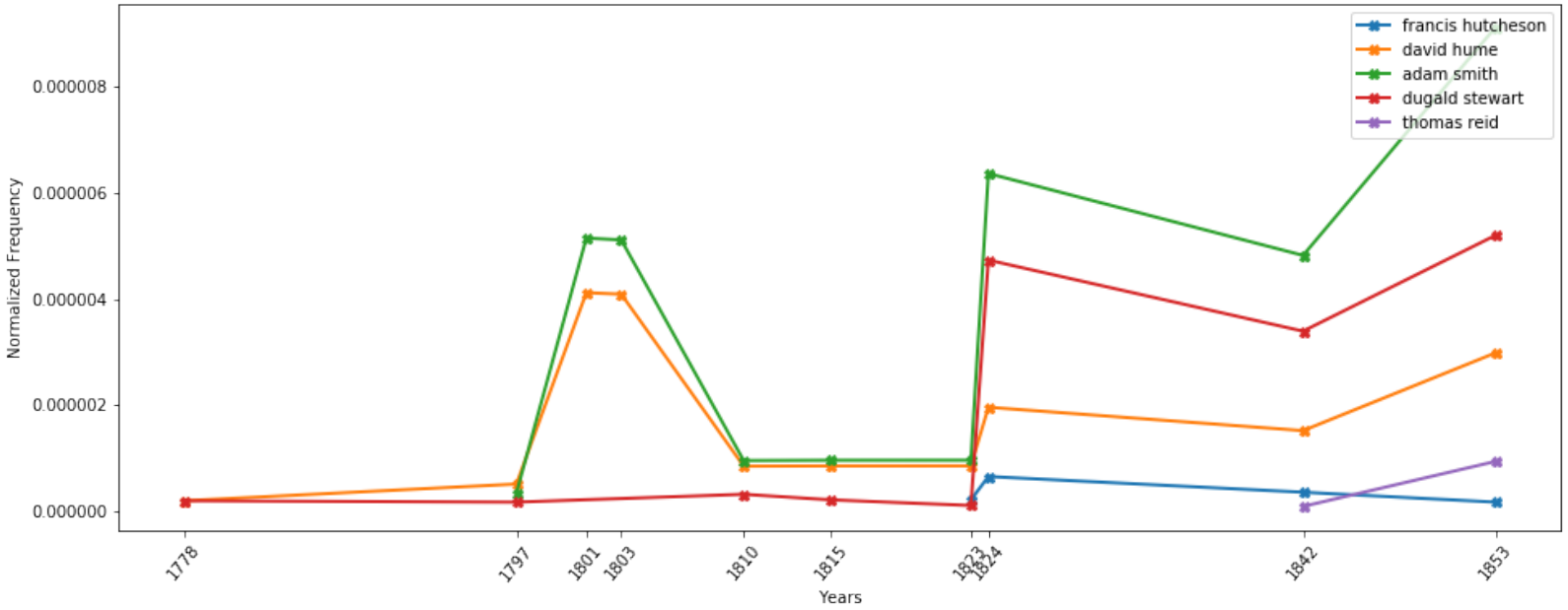
Gazetteers of Scotland

Ladies' Edinburgh Debating Society

https://data.nls.uk/data/digitised-collections/a-medical-history-of-british-india/

National Library Scottish – Encyclopaedia Britannica

Exploring the popularity of Scottish Philosophers over time



N-gram for the normalised frequencies of the Scottish philosophers

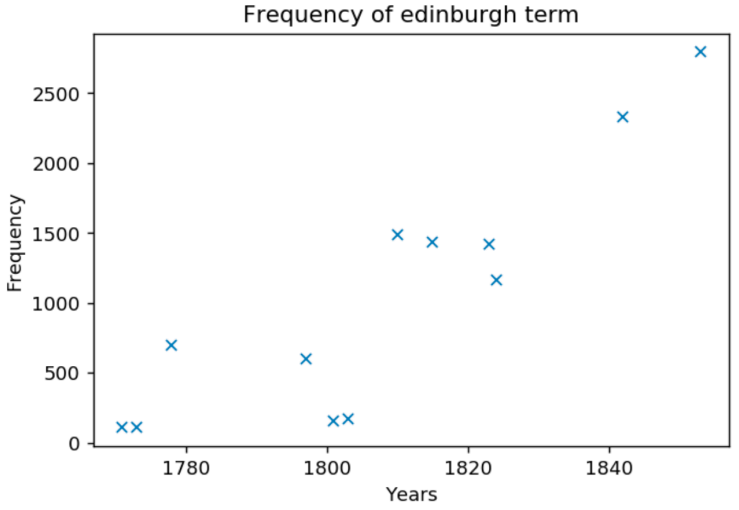
Check our Jupyter Notebooks

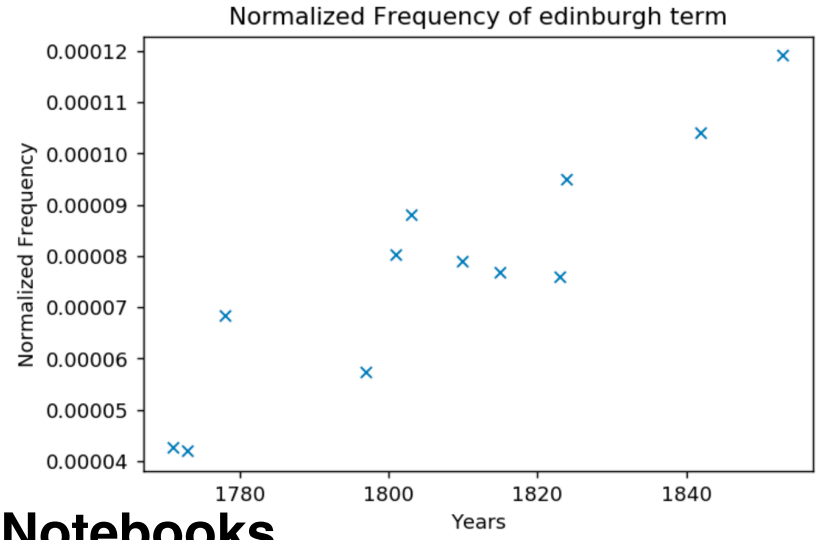
https://github.com/alan-turing-institute/defoe_visualization/tree/master/NLS

National Library Scottish – Encyclopaedia Britannica

Exploring the term “Edinburgh ”

	year	words_left	term	words_right
25	1797	from', 'the', 'accounts', 'given', 'by', 'the', 'late', 'dr', 'whytt', 'of'	edinburgh	'and', 'others', 'it', 'should', 'appear', 'that', 'this', 'had', 'sometimes', 'happe...
34	1797	stirling', 'and', 'frequently', 'the', 'three', 'principal', 'fortrefles', 'of', 'the'...	edinburgh	'stirling', 'and', 'dunbarton', 'the', 'last', 'heir', 'of', 'the', 'scottish', 'mona...
35	1797	regent', 'mar', 'after', 'ihe', 'was', 'obliged', 'by', 'the', 'treaty', 'of'	edinburgh	'to', 'desist', 'from', 'wearing', 'the', 'arms', 'of', 'england', 'in', 'the']
94	1797	considered', 'as', 'presages', 'of', 'a', 'pestilence', 'they', 'appear', 'annually', ...	edinburgh	'in', 'february', 'and', 'feedon', 'the', 'berries', 'of', 'the', 'mountainalh', 'they']
350	1797	into', 'inches', 'the', 'flandard', 'is', 'kept', 'in', 'the', 'couneilehamber', 'of'	edinburgh	'and', 'being', 'compared', 'with', 'the', 'englilh', 'yard', 'is', 'found', 'to']
437	1797	with', 'in', 'duddingilonloch', 'a', 'freshwater', 'lake', 'within', 'a', 'mile', 'of'	edinburgh	'in', 'france', 'it', 'stays', 'throughout', 'the', 'year', 'and', 'makes', 'a']
448	1797	comprehensive', 'system', 'in', 'three', 'volumes', 'publiffied', 'by', 'mr', 'elliot'...	edinburgh	'upon', 'a', 'plan', 'approved', 'of', 'by', 'dr', 'monro', 'and', 'executed']
487	1797	and', 'on', 'the', 'th', 'of', 'june', 'held', 'a', 'parliament', 'at'	edinburgh	'at', 'that', 'time', 'the', 'clergy', 'met', 'in', 'one', 'of', 'the']
488	1797	mention', 'is', 'made', 'of', 'his', 'being', 'minister', 'of', 'pencaitland', 'near'	edinburgh	'in', 'but', 'we', 'find', 'nothing', 'said', 'there', 'or', 'any', 'where']
525	1797	mild', 'difpolitions', 'and', 'peaceable', 'deportment', 'a', 'parliament', 'was', 'su...	edinburgh	'in', 'summer', 'and', 'the', 'duke', 'was', 'appointed', 'commiffioner', 'besides', ...





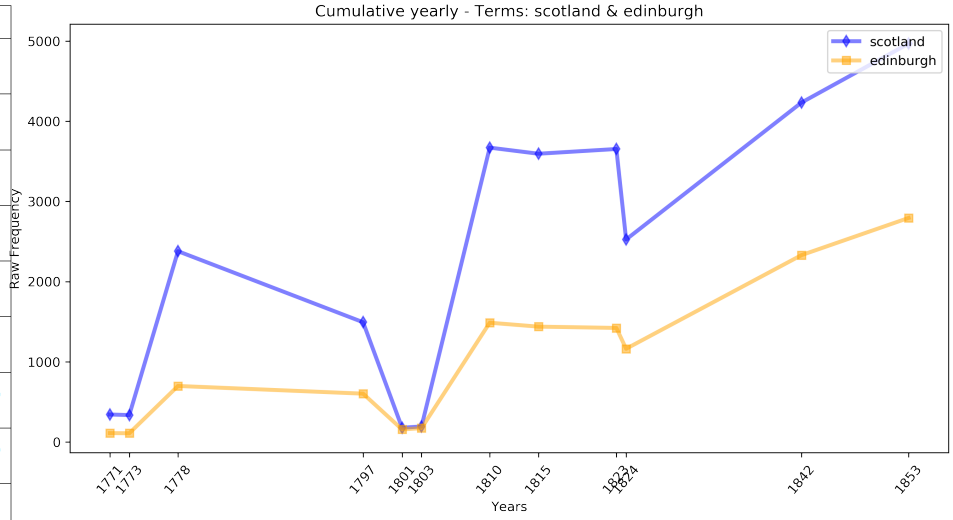
Check our Jupyter Notebooks

https://github.com/alan-turing-institute/defoe_visualization/tree/master/NLS

National Library Scottish – Encyclopaedia Britannica

Exploring the term “Edinburgh ”

	year	words_left	term	words_right
25	1797	from', 'the', 'accounts', 'given', 'by', 'the', 'late', 'dr', 'whytt', 'of'	edinburgh	'and', 'others', 'it', 'should', 'appear', 'that', 'this', 'had', 'sometimes', 'happe...
34	1797	stirling', 'and', 'frequently', 'the', 'three', 'principal', 'fortrefles', 'of', 'the'...	edinburgh	'stirling', 'and', 'dunbarton', 'the', 'last', 'heir', 'of', 'the', 'scottish', 'mona...
35	1797	regent', 'mar', 'after', 'ihe', 'was', 'obliged', 'by', 'the', 'treaty', 'of'	edinburgh	'to', 'desist', 'from', 'wearing', 'the', 'arms', 'of', 'england', 'in', 'the']
94	1797	considered', 'as', 'presages', 'of', 'a', 'pestilence', 'they', 'appear', 'annually', ...	edinburgh	'in', 'february', 'and', 'feedon', 'the', 'berries', 'of', 'the', 'mountainalh', 'they']
350	1797	into', 'inches', 'the', 'flandard', 'is', 'kept', 'in', 'the', 'couneilehamber', 'of'	edinburgh	'and', 'being', 'compared', 'with', 'the', 'englilh', 'yard', 'is', 'found', 'to']
437	1797	with', 'in', 'duddingilonloch', 'a', 'freshwater', 'lake', 'within', 'a', 'mile', 'of'	edinburgh	'in', 'france', 'it', 'stays', 'throughout', 'the', 'year', 'and', 'makes', 'a']
448	1797	comprehensive', 'system', 'in', 'three', 'volumes', 'publiffied', 'by', 'mr', 'elliot'...	edinburgh	'upon', 'a', 'plan', 'approved', 'of', 'by', 'dr', 'monro', 'and', 'executed']
487	1797	and', 'on', 'the', 'th', 'of', 'june', 'held', 'a', 'parliament', 'at'	edinburgh	'at', 'that', 'time', 'the', 'clergy', 'met', 'in', 'one', 'of', 'the']
488	1797	mention', 'is', 'made', 'of', 'his', 'being', 'minister', 'of', 'pencaitland', 'near'	edinburgh	'in', 'but', 'we', 'find', 'nothing', 'said', 'there', 'or', 'any', 'where']
525	1797	mild', 'difpolitions', 'and', 'peaceable', 'deportment', 'a', 'parliament', 'was', 'su...	edinburgh	'in', 'summer', 'and', 'the', 'duke', 'was', 'appointed', 'commiffioner', 'besides', ...



Check our Jupyter Notebooks

https://github.com/alan-turing-institute/defoe_visualization/tree/master/NLS

National Library Scottish – Encyclopaedia Britannica


```
"Encyclopaedia Britannica: or, A dictionary of arts and sciences':
- archive_name: /home/tdm/datasets/eb_test/144850366
  articles:
    ACQUEST: or Acquist, in law, signifies goods got by purchase or donation. See
      CoNtiUEST.
    ACQUI: "a town of Italy, in the Dutchy of Montferrat, with a biihop\u2019s see,\
      \ and commodious baths. It was taken by the Spaniards in 1745, and retaken by\
      \ the Piedmontese in 1746; but after this, it was taken again and difrcantled\
      \ by the French, who afterwards forsook it. It is seated on the river Bormio,\
      \ 25 milesN.W. Genoa, and 30S. Cafal, 8. 30.E. 44. 40. lat."
    ACQUIESCENCE: in commerce, is the consent that a person gives to the determination
      given either by arbitration, orbyaconful
    ACQUIETANDIS plegiis: in the English law, is a writ that lies for a surety, againd
      a creditor, who refufes to acquit the complainant after the debt is paid.
    ACQUIETANTIA de Jhiris et hundredh: in England, signifies the privilege of being
      free from suit and service in fliires and hundreds.
    ACQUISITION: in general, denotes the obtaining or procuring something. Among lawyers,
      it is used for the right or title to an estate got by purchase or dona
    ACQUITARE: in ancient law-books, signifies to difcliarge or pay off the debts
      of a person deceased.
    ACQUITTAL: a discharge, deliverance, or setting of a person free from the guilt
      or suspicion of an offence.
    ACQUITTANCE: a release or discharge in writing for a sum of money.
    ACRA: a town of Africa, on the coast of Guinea, where the English, Dutch, and
      Danes, have strong sorts, and each sort its particular village, o. 2.W. 5. o.
      lat.
    ACRASIA: among physicians, signifies the predominancy of one quality over another.
    ACRE: "or Acra, a sea-port town in Syria. It was formerly called Ptolemais, and\
      \ is a bishop\u2019s see. It was very famous in the time of the crusadoes, and\
      \ underwent several sieges both by the Chrif Hans and Saracens. It is now an\
      \ inconsiderable town, being entirely supported by its harbour, which, is frequented\
      \ by (hips of several nations. It is 20 milesS. Tyre, and 37N. Jerusalem, 39.\
      \ 25.E. 32. 40. lat. Acre, in the Mogul\u2019s dominions, the same with lack,\
      \ and signifies the sum of 100,000 rupees ; the rupee is -of the value of the\
      \ French crown of 3 livres, or 30 fols of Holland; an 100 lacks of rupees make\
      \ a couron in Indoftan, or 10,000,000 rupees; the pound Sterling is about 8\
      \ rupees; according to which proportion, a lack of rupees amounts to 12,500\
      \ pounds Sterling. Acre, a measure of land used in several provinces of France,\
      \ particularly in Normandy. It is larger or less according to the different\
      \ places ; but commonly contains 160 perches. The Acre of woods in France, consists\
      \ of four roods, called verg es; the rood is 40 perches, the perch 24 feet,\
      \ the foot 12 inches, the inch 12 lines. Acre, the universal measure of land\
      \ in Britain. An acre in England contains 4 square roods, a rood 40ACR or poles\
      \ of i6\xA3 feet each byftatute. Yet this measure does not prevail in all parts\
      \ of England, as the length of the pole varies in different counties, and is\
      \ called cujiosnary measure, the difference running from the 16feet to 28. The\
      \ acre is also divided into xo square chains, of 22 yards each, that is 4840\
      \ square yards. An acre in Scotland contains 4 square roods ; 1 square rood\
      \ is 40 square falls; 1 square fall, 36 square ells; 1 square ell, 9 square\
      \ feet, and 73 square inches; 1 square foot, 144 square inches. The Scots acre\
      \ is also divided into 10 square chains ; the measuring chr.in Ihould be 24\
      \ ells in length, divided into 100 links, each link 8 r \u2019 c 7I and soI\
      \ chain will contain 10,000 square links. The Englilh (latute acre is about\
      \ 3 roods and 6 falls standard measure of Scotland."
    ACREME: in old law-books, signifies ten acres of land.
    ACRIBEIA: signifies great accuracy.
    ACRID: a name for any thing that is of a (harp or pungent taste.
    ACRIDOPHAGI: "signifies locujl-eaters. It has been much disputed whether the inhabitants\
      \ of Arabia, Ethiopia, <bc ever eat locusts. We Ihall give the substance of\
      \ what Haffelquill says on this fubjedt, who travelled in Syria and Egypt so\
      \ late as the year 1752. This ingenious gentleman, who travelled with a view\
      \ to improve natural history, informs us, that he asked Franks, and many other\
      \ people who had lived long in these countries, whether they had ever heard\
      \ that the inhabitants of Arabia and Ethiopia, &c. used locusts as food. They\
      \ answered that they had. He likewise alked the same question of Armenians,\
      \ Cophtes, and Syrians, who lived in Arabia, and had travelled in Syria and\
      \ near the Red-sea; some of whom said they heard of such a practice, and others\
      \ that they, had often seen the people eat these infedts. He at last obtained\
      \ complete fatisfadtion on this head from a learned Iheck at Cairo, who had\
      \ lived six years in Mecca. This gentleman told him, in presence ofM. Grand,\
      \ the principal French interpreter at Cairo, and others, that a famine frequently\
      \ rages at Mecca when there is a scarcity of corn in Egypt, which obliges the\
      \ inhabitants to live upon coarser food than ordinary : That when corn is scarce,\
      \ the Arabians grind the loculls in hand-mills, or Hone mortars, and bake them\
      \ into cakes, and use these cakes in place of bread: That he has frequently\
      \ seen locusts used by the Arabians, even when there was no scarcity of corn;\
      \ but then they boil them, itew them with butter, and make them into a kind\
      \ of fricaffee, which he says is not disagreeably tailed; for he had sometimes\
      \ tailed these locuft-fricaffees out of curiosity. From this account, we may\
      \ see the folly of that dispute among divines about the nature of St John\u2019\
      s food in the wilderness. Some of them say that loculls were the fruits of certain\
      \ trees, others that they were a kind of birds, <bc.; but those who adhered\
```


ACQUEST, or ACQUIST, in law, fignifies goods got by purchafe or donation. See CONQUEST.

ACQUI, a town of Italy, in the Dutchy of Montferrat, with a bifhop's fee, and commodious baths. It was taken by the Spaniards in 1745, and retaken by the Piedmonteſe in 1746; but after this, it was taken again and difmantled by the French, who afterwards forſook it. It is ſeated on the river Bormio, 25 miles N. W. of Genoa, and 30 S. of Cafal, 8. 30. E. long. 44. 40. lat.

ACQUIESCENCE, in commerce, is the conſent that a perſon gives to the determination given either by arbitration, or by a conſul.

ACQUIETANDIS *plegiis*, in the Engliſh law, is a writ that lies for a ſurety, againſt a creditor, who refuſes to acquit the complainant after the debt is paid.

ACQUIETANTIA *de ſhirit et hundredis*, in England, fignifies the privilege of being free from ſuit and ſervice in ſhires and hundreds.

ACQUISITION, in general, denotes the obtaining or procuring ſomething. Among lawyers, it is uſed for the right or title to an eſtate got by purchaſe or donation.

ACQUITARE, in ancient law-books, fignifies to diſcharge or pay off the debts of a perſon deceaſed.

ACQUITTAL, a diſcharge, deliverance, or ſetting of a perſon free from the guilt or ſuſpicion of an offence.

ACQUITTANCE, a releaſe or diſcharge in writing for a ſum of money.

ACRA, a town of Africa, on the coaſt of Guinea, where the Engliſh, Dutch, and Danes, have ſtrong forts, and each fort its particular village, o. 2. W. long. 5. o. lat.

ACRASIA, among phyſicians, ſignifies the predominancy of one quality over another.

ACRE, or ACRA, a ſea-port town in Syria. It was formerly called *Ptolemais*, and is a biſhop's ſee. It was very famous in the time of the cruſadoes, and underwent ſeveral ſieges both by the Chriſtians and Saracens. It is now an inconſiderable town, being entirely ſupported by its harbour, which is frequented by ſhips of ſeveral nations. It is 20 miles S. of Tyre, and 37 N. of Jeruſalem, 39. 25. E. long. 32. 40. lat.

ACRE, in the Mogul's dominions, the ſame with lack, and ſignifies the ſum of 100,000 rupees; the rupee is of the value of the French crown of 3 livres, or 30 ſols of Holland; an 100 lacks of rupees make a couron in Indoſtan, or 10,000,000 rupees; the pound Sterling is about 8 rupees; according to which proportion, a lack of rupees amounts to 12,500 pounds Sterling.

ACRE, a meaſure of land uſed in ſeveral provinces of France, particularly in Normandy. It is larger or leſs according to the different places; but commonly contains 160 perches.

The ACRE *of woods in France*, conſiſts of four roods, called *verg es*; the rood is 40 perches, the perch 24 feet, the foot 12 inches, the inch 12 lines.

ACRE, the univerſal meaſure of land in Britain. An acre in England contains 4 ſquare roods, a rood 40 perches or poles of 16½ feet each by ſtatute. Yet this meaſure does not prevail in all parts of England, as the length of the pole varies in different counties, and is called *cuſtomary meaſure*, the difference running from the 16½ feet to 28. The acre is alſo divided into 10 ſquare chains, of 22 yards each, that is 4840 ſquare yards. An acre in Scotland contains 4 ſquare roods; 1 ſquare rood is 40 ſquare falls; 1 ſquare fall, 36 ſquare ells; 1 ſquare ell, 9 ſquare feet, and 73 ſquare inches; 1 ſquare foot, 144 ſquare inches. The Scots acre is alſo divided into 10 ſquare chains; the meaſuring chain ſhould be 24 ells in length, divided into 100 links, each link 8$\frac{7}{10}$ inches; and ſo 1 ſquare chain will contain 10,000 ſquare links.

The Engliſh ſtatute acre is about 3 roods and 6 falls ſtandard meaſure of Scotland.

ACREME, in old law-books, ſignifies ten acres of land.

ACRIBEIA, ſignifies great accuracy.

ACRID, a name for any thing that is of a ſharp or pungent taſte.

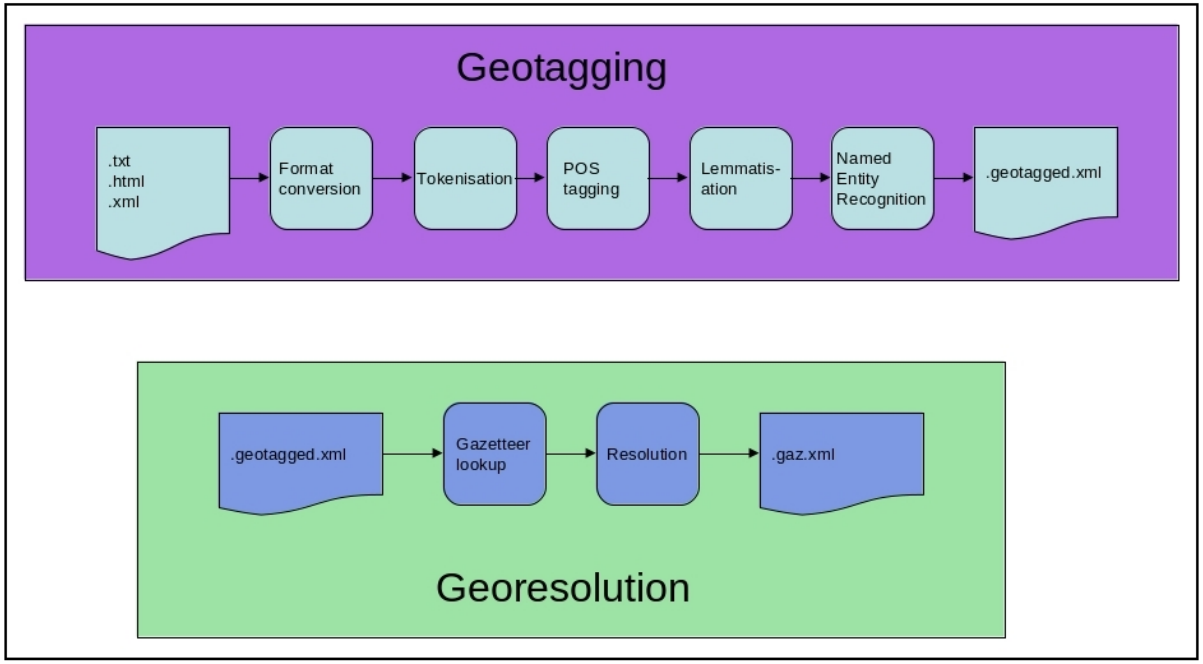
ACRIDOPHAGI, ſignifies *locuſt-eaters*. It has been much diſputed whether the inhabitants of Arabia, Ethiopia, *&c.* ever eat locuſts. We ſhall give the ſubſtance of what Haſſelquiſt ſays on this ſubject, who travelled in Syria and Egypt ſo late as the year 1752. This ingenious gentleman, who travelled with a view to improve natural hiſtory, informs us, that he aſked Franks, and many other people who had lived long in theſe countries, whether they had ever heard that the inhabitants of Arabia and Ethiopia, *&c.* uſed locuſts as food. They anſwered that they had. He likewiſe aſked the ſame queſtion of Armenians, Cophtes, and Syrians, who lived in Arabia, and had travelled in Syria and near the Red-ſea; ſome of whom ſaid they heard of ſuch a practice, and others that they had often ſeen the people eat theſe inſects. He at laſt obtained complete ſatisfaction on this head from a learned ſheck at Cairo, who had lived ſix years in Mecca. This gentleman told him, in preſence of M. le Grand, the principal French interpreter at Cairo, and others, that a famine frequently rages at Mecca when there is a ſcarcity of corn in Egypt, which obliges the inhabitants to live upon coarſer food than ordinary: That when corn is ſcarce, the Arabians grind the locuſts in hand-mills, or ſtone mortars, and bake them into cakes, and uſe theſe cakes in place of bread: That he has frequently ſeen locuſts uſed by the Arabians, even when there was no ſcarcity of corn; but then they boil them, ſtew them with butter, and make them into a kind of fricaſſee, which he ſays is not diſagreeably taſted; for he had ſometimes taſted theſe locuſt-fricaſſees out of curioſity. From this account, we may ſee the folly of that diſpute among divines about the nature of St John's food in the wilderneſs. Some of them ſay that locuſts were the fruits of certain trees, others that they were a kind of birds, *&c.*; but thoſe who adhered to the literal meaning of the text were at leaſt the moſt orthodox, although their arguments were perhaps not ſo ſtrong as they might have been, had they had an opportunity of quoting ſuch an author as Haſſelquiſt.

ACRI-

National Library Scottish – Scottish Gazetteers

Geoparsing the historical Gazetteers of Scotland

Edinburgh Geoparser + Defoe



Edinburgh Geoparser Pipeline

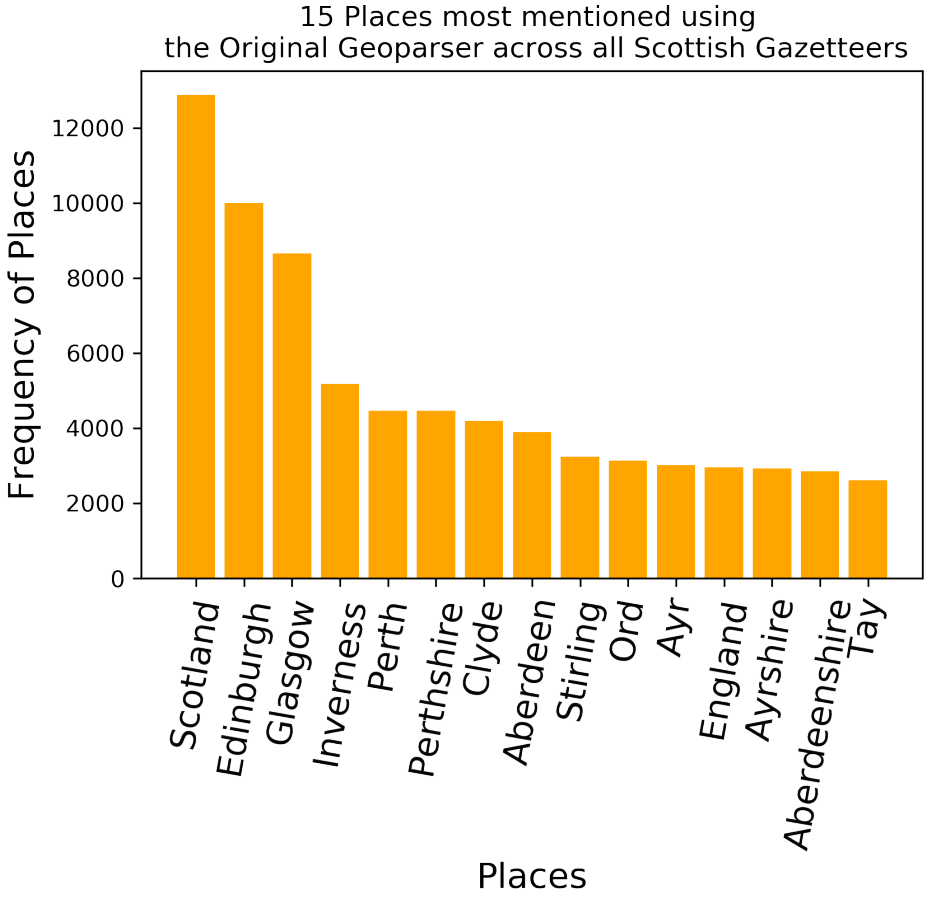
Process input text to identify and classify place-name entities

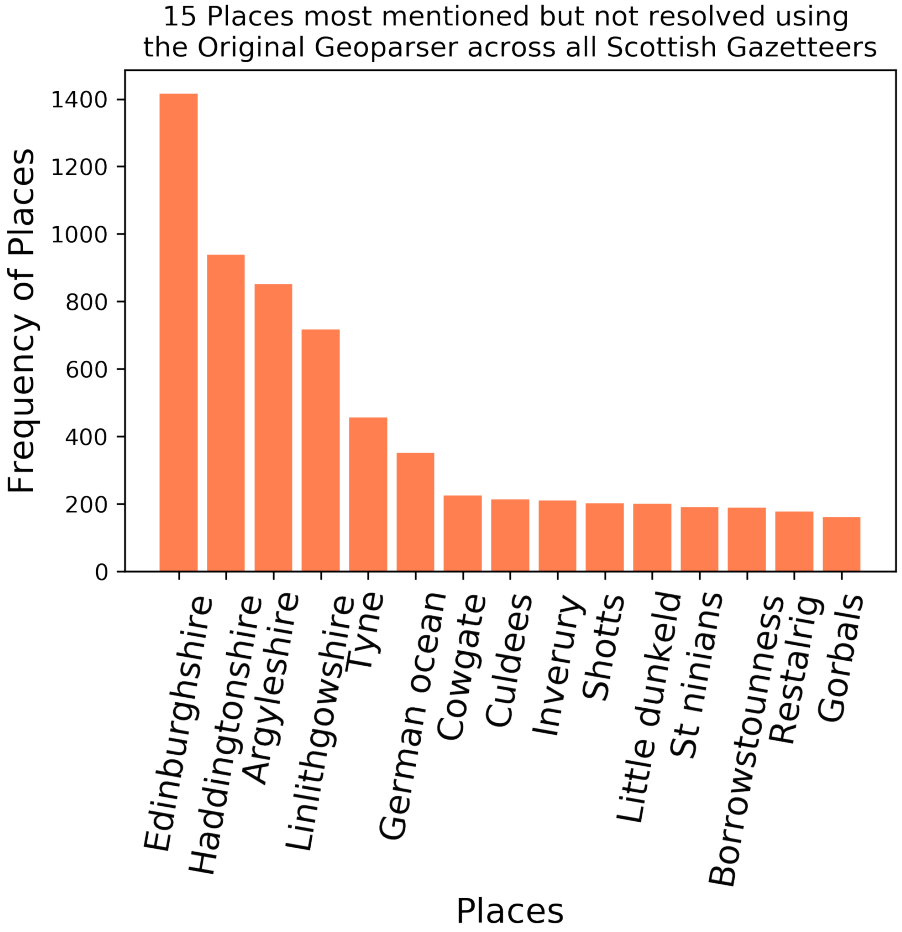
Get the coordinates off place-name entities – uses an online gazetteer (e.g. *geonames*)

Collaboration with the **Language Technology Group** – University of Edinburgh

https://www.ltg.ed.ac.uk/software/geoparser/

National Library Scottish – Scottish Gazetteers

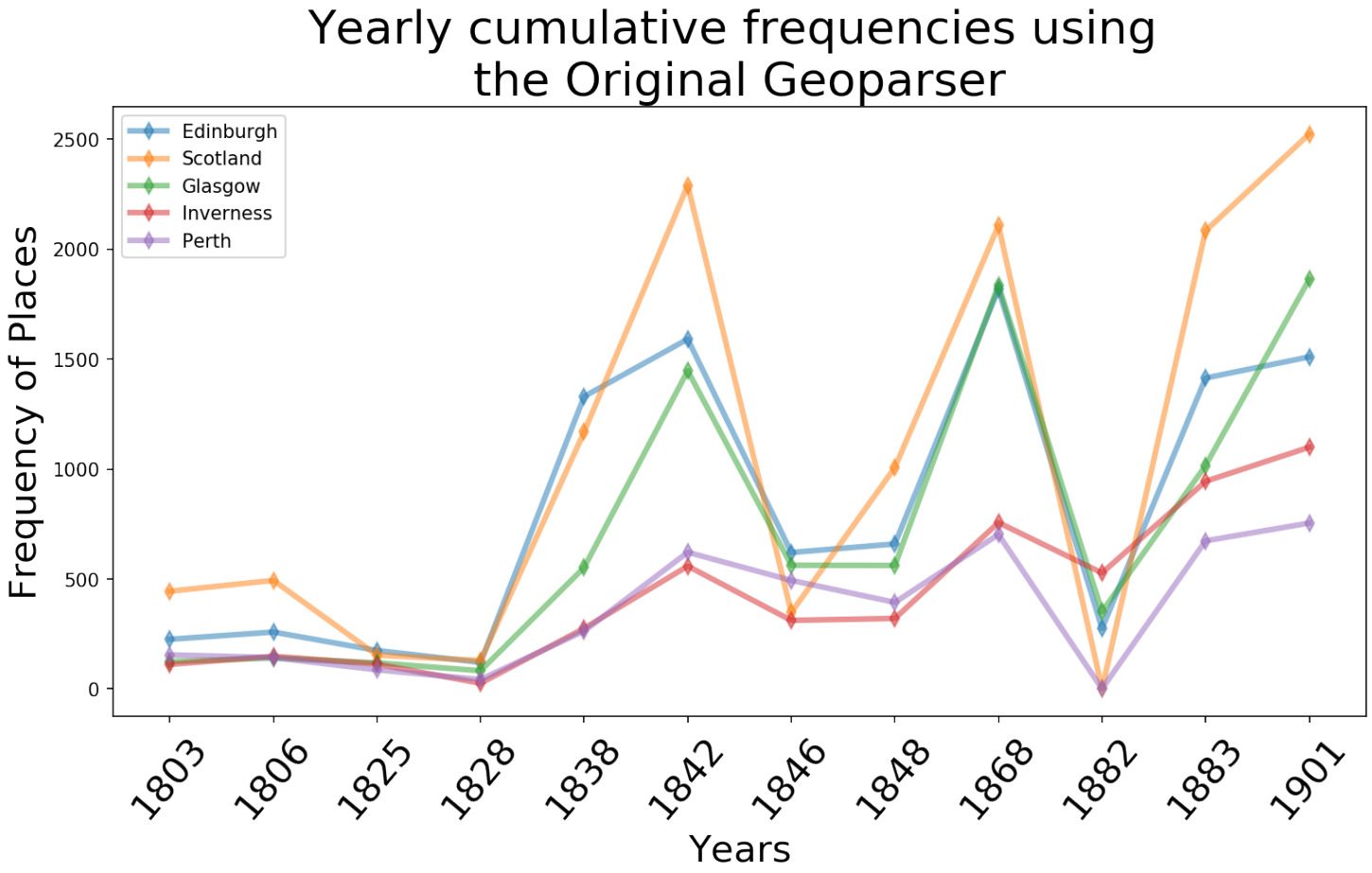




Check our Jupyter Notebooks

https://github.com/alan-turing-institute/defoe_visualization/blob/master/Scottish_Gazetteer/

National Library Scottish – Scottish Gazetteers



Publication: Geoparsing the Historical Gazetteers of Scotland: Accurately Computing Location in Mass Digitised Texts, 8th Workshop on the Challenges in the Management of Large Corpora, 16th May 2020

CDCS TEXT MINING LAB – CALL FOR PROJECTS

The Centre for Data, Culture & Society is looking for humanities and social science researchers who can ask complex questions of large-scale data sets.

Deadline for expressions of interest: April 20th - Applications are now closed.

This call is open to research active staff and PhD students within CAHSS at the University of Edinburgh.

****Please note that we envisage being able to conduct this activity remotely.****

NTRODUCING DEFOE

CDCS TEXT MINING LAB – CALL FOR PROJECTS

DATA SETS

- **BRITISH LIBRARY BOOKS**

Over 68,000 books from the 16th to the 19th century, covering geography, philosophy, history, poetry and literature in a variety of languages.

- **BRITISH LIBRARY NEWSPAPERS**

1TB of digitised British newspapers from the 18th to the early 20th Century

- **TIMES DIGITAL ARCHIVE**

All the articles in 69,699 volumes of The Times newspaper between 1785 and 2009.

- **PAPERS PAST: NEW ZEALAND AND PACIFIC NEWSPAPERS**

Over 5 million pages of New Zealand and Pacific newspapers from the 19th and 20th Centuries.

- **GAZETTEERS OF SCOTLAND**

20 volumes of the most popular 19th Century gazetteers of Scotland, including detailed historical and geographic information about each place.

- **ENCYCLOPAEDIA BRITANNICA 1768 - 1860**

The first 8 volumes of the Encyclopaedia Britannica, issued from 1768-1860, comprising a total of 143 volumes. 155,388 pages, 166m words.

***Coming soon*:**

- JISC MEDICAL HERITAGE LIBRARY
- HANSARD ARCHIVE
- STATISTICAL ACCOUNTS OF SCOTLAND
- DIGITISED THESES FROM EDINBURGH UNIVERSITY LIBRARY

If we don't have what you're looking for in term of data sets, you can propose a new data set for use with *defoe*.

Conclusions

- New digital toolbox for mining historical data.
- Text analyses across large collections in parallel.
- Rich set of text mining queries.
- NLP prepossessing techniques to mitigate OCR errors.
- Portability on different computing environments
- Compatibility with different storage systems
- Integration with other NLP tools – e.g. Edinburgh Geoparser

”All this work provides the means to search across large scale datasets and to return results for further analysis and interpretation by historians.”

Analysing Digital Historical Textual Data at Scale with Defoe Toolbox

Dr. Rosa Filgueira, EPCC,
University of Edinburgh
Email: rosa.filgueira@ed.ac.uk