Defoe: A Spark-based for Analysing Digital Historical Textual Data

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

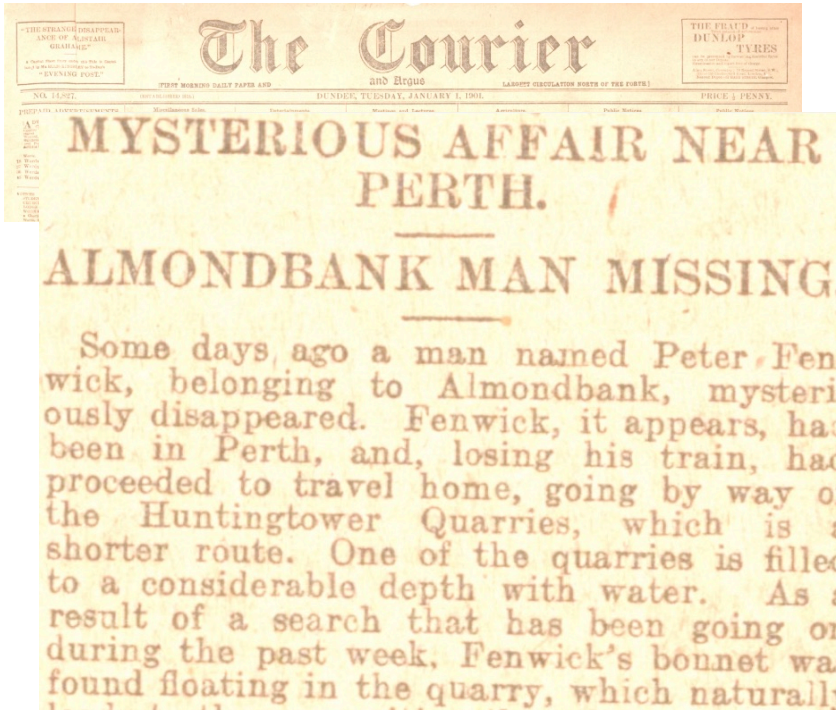
Funding
- ATI–SE Data Engineering Programme
- Living with Machines (LwM)
- Text Data Mining (TDM) (staring next month)

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

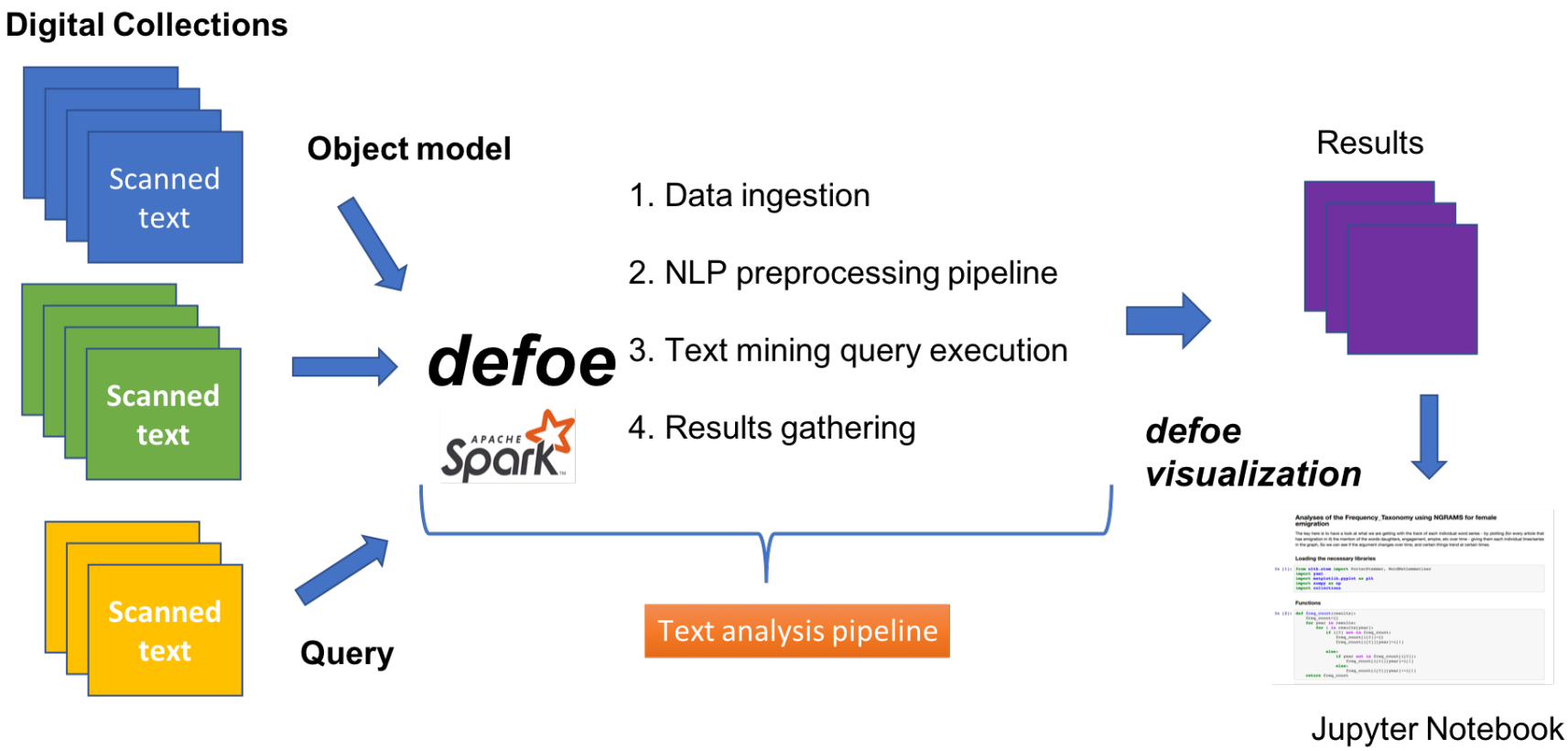
MYSTERIOUS AFFAIR NEAR PERTH.

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

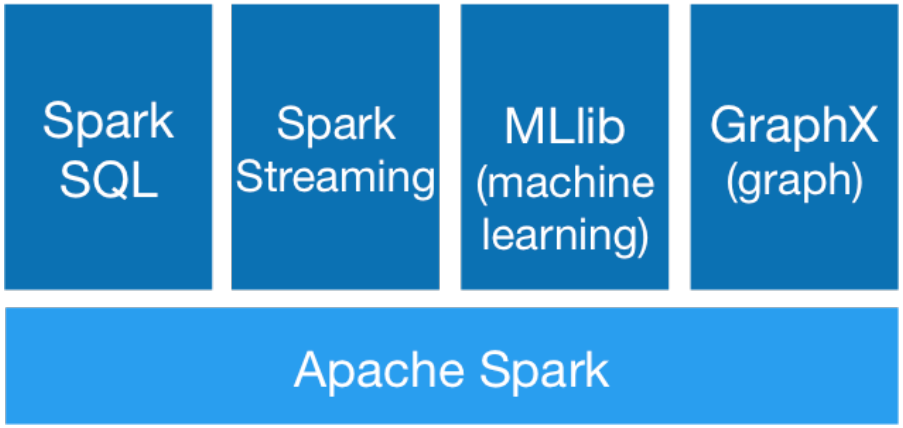


scalable portable text mining queries large dataset

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

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

Apache Spark



https://spark.apache.org/

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

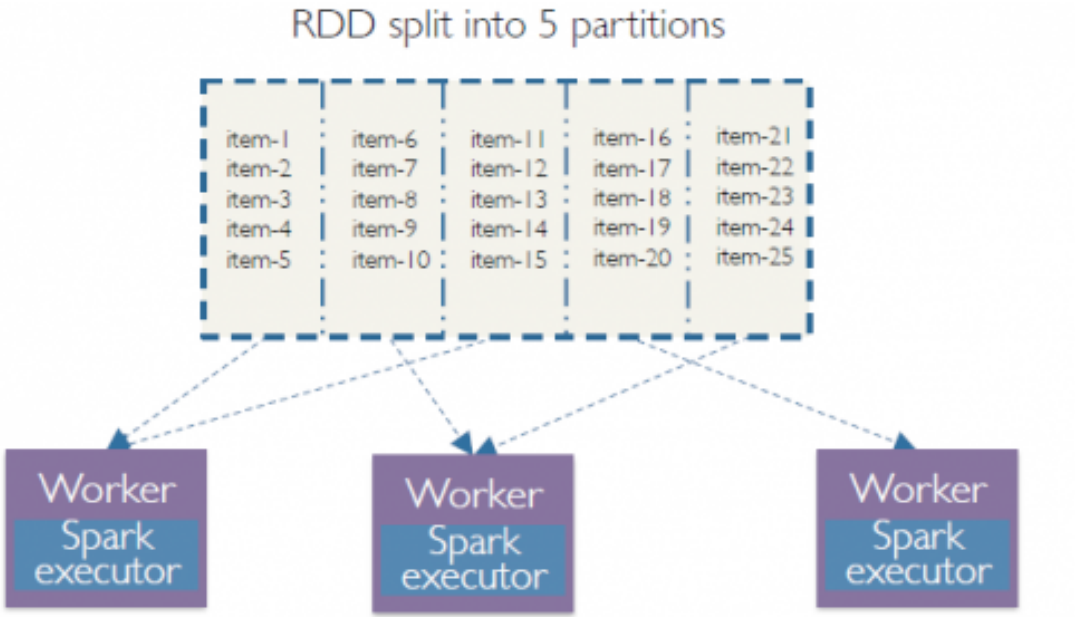
  Java, Scala, **Python** and R

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

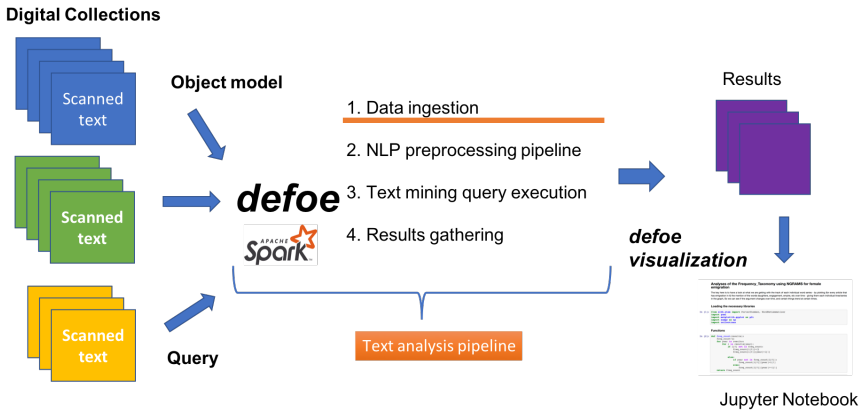
Apache Spark

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



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 metadata document and a XML per page

Object models -- loading data into RDDs:

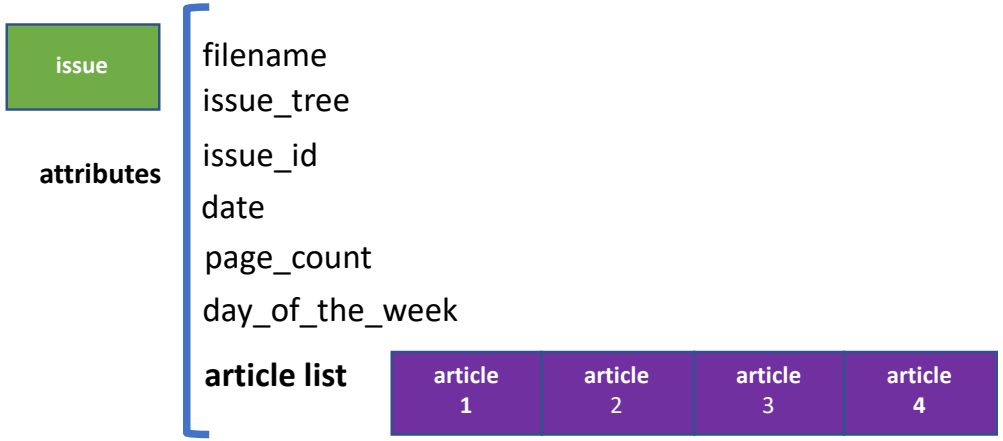
PAPER (physical rep num1)
NZPP (physical rep num2)
ALTO (physical rep num3)
FMP (physical rep num3)

PAPER object model (British Library Newspapers)

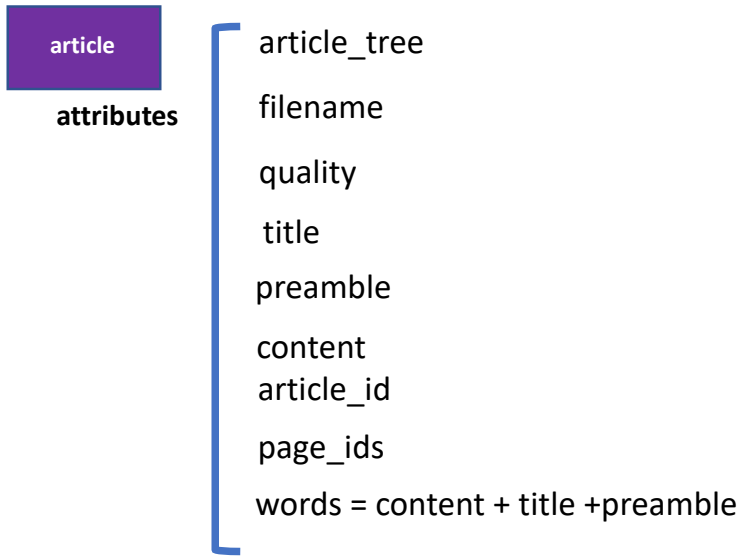
0000164- The Courier and Argus
0000187- The Bath Chronicle
0000195- Archer Bath Chronicle
0000321- The Nottingham Evening Post
0000452- Edinburgh Evening News

RDDs

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



Class Article → Representation of an article

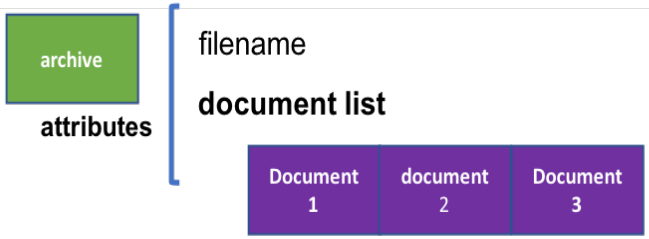


ALTO object model (British Library books)

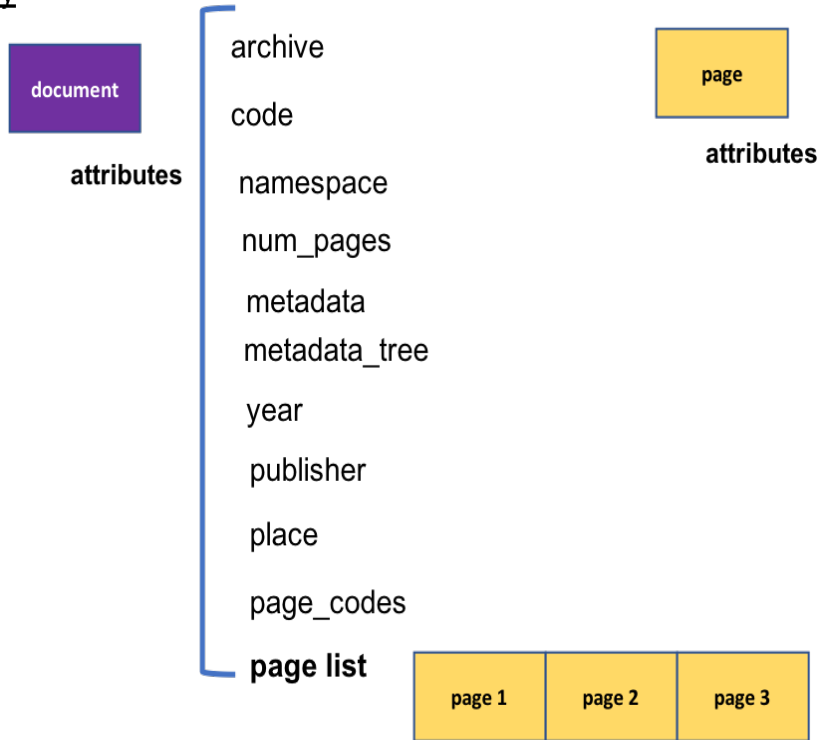
RDDs

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

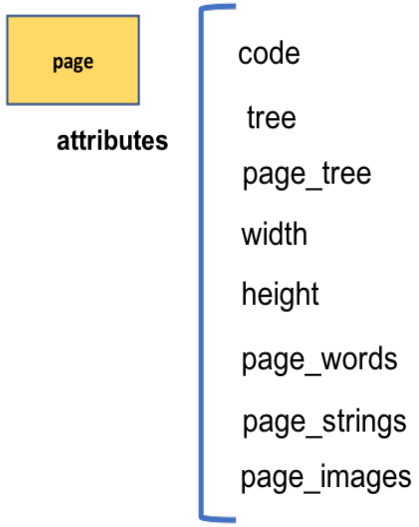
Class Archive → Representation of a zipped archive



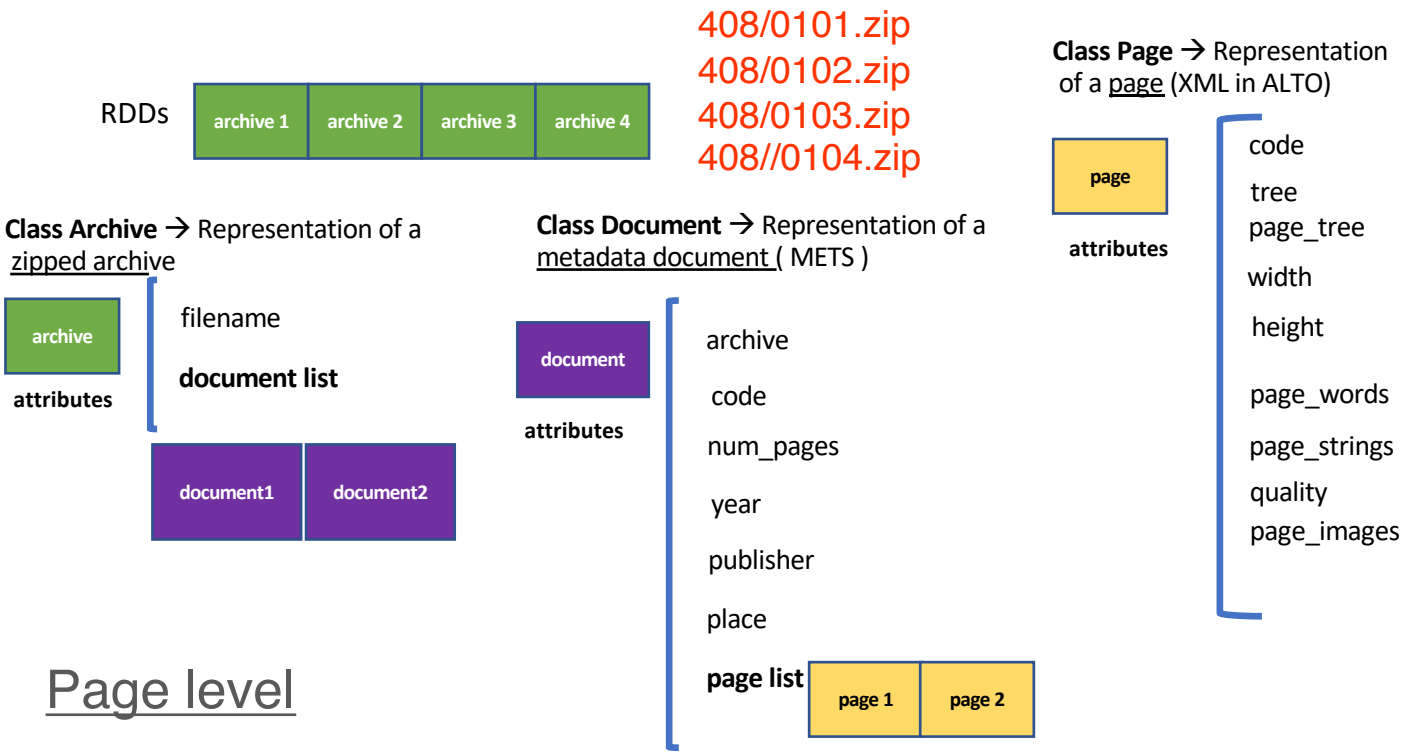
Class Document → Representation of a metadata document (XML in METS/MODS) and an ALTO directory



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

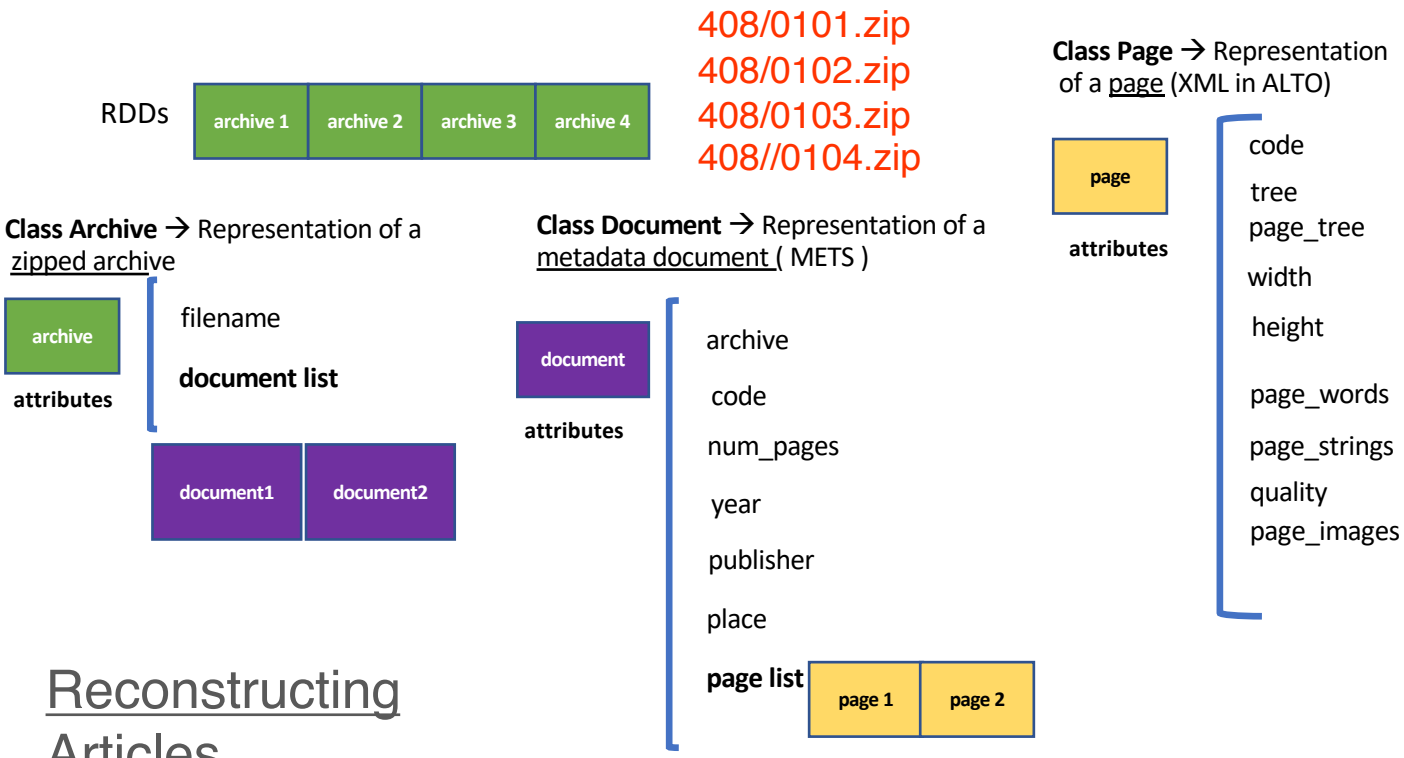


FMP object model (Find My Past Newspapers)



https://www.brodrigues.co/blog/2019-01-13-newspapers_mets_alto/

FMP object model (Find My Past Newspapers)



Reconstructing Articles

METS

logical structure - how many articles has an issue, and each article id (*art0001, art0002, art00003*)

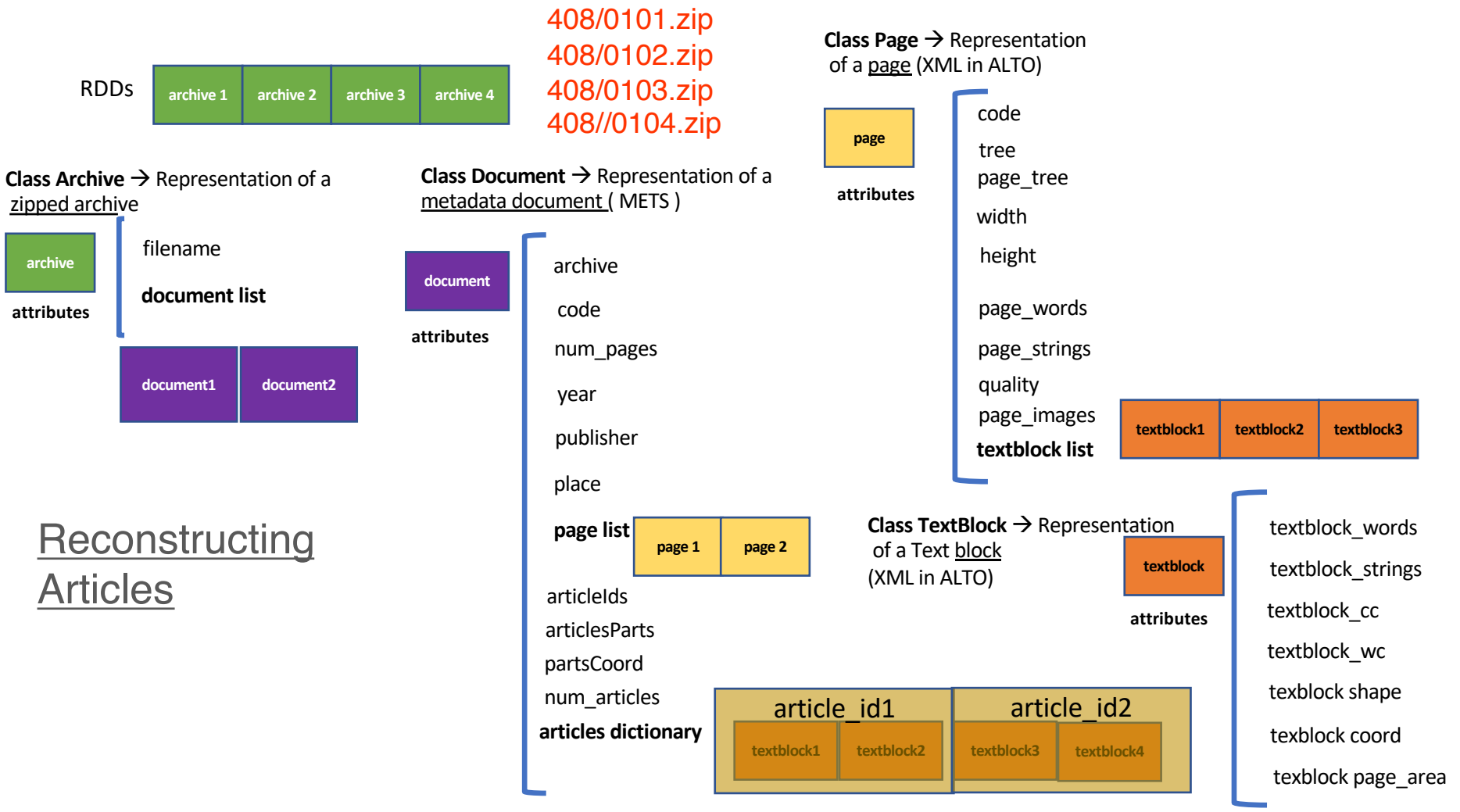
structural link - how many text blocks/parts has each article, and each text blocks id *(pa0001001, pa0001002,).*

physical structure - coordinates of each text block (*COORDS="1220,5,2893,221"*)

ALTO

Text blocks/parts ids that has each page and their content

FMP object model (Find My Past Newspapers)



Reconstructing Articles

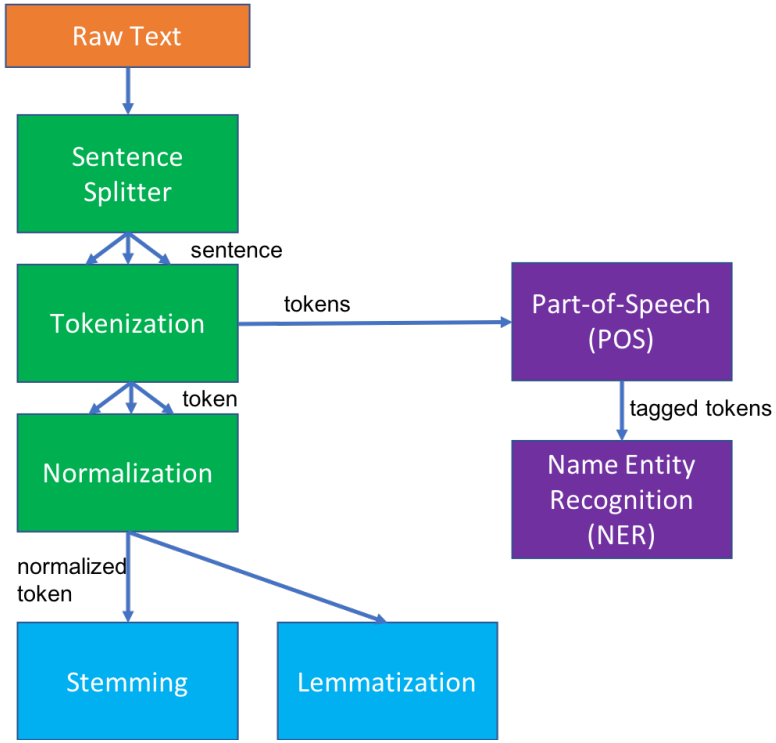
(It also works with unzipped archives)

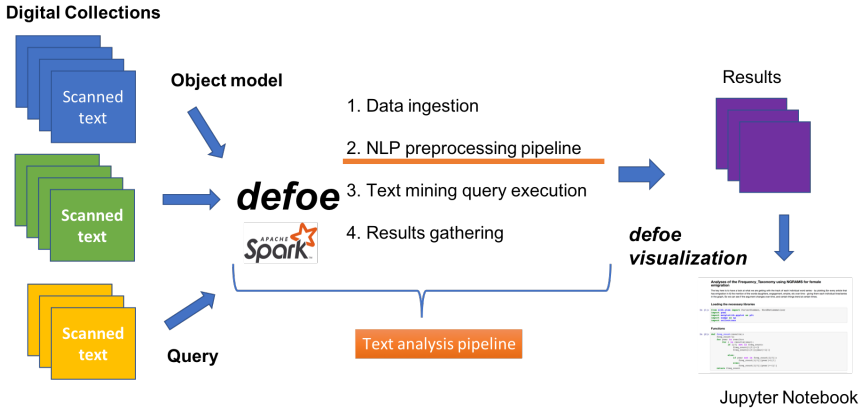
Digital Collections

Important

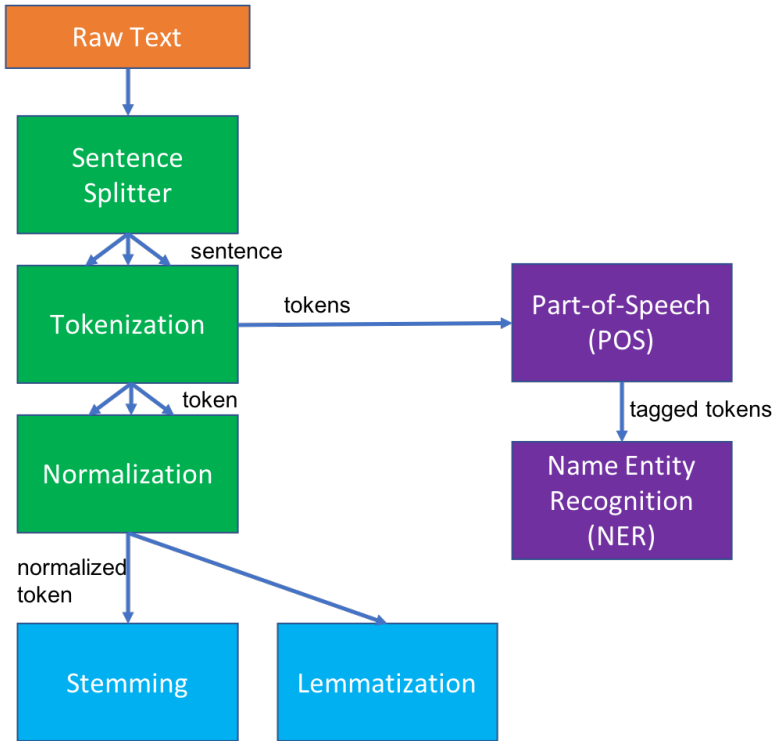
	Dataset	Period	Structure	XML Schema	Space	Model
ATI-SE	**British Library Books (BLB)**	1510-1899	ZIP per book - XML metadata - XML per page	METS and ALTO schemas	~220GB	**ALTO**
ATI-SE	**British Library Newspapers (BLN)**	1714-1950	XML per issue	GALEN Schema	~1TB	**PAPERS**
ATI-SE	**Times Digital Archive (TDA)**	1785-2009	XML per issue	GALEN Schema	~324GB	**PAPERS**
ATI-SE	**Papers Past New Zealand and Pacific newspapers (NZPP)**	1839-1863	XML per 22 articles	XML from a search via an API	~4GB	**NZPP**
LwM	**FindMyPast (FPM)**	1752 - 1957	-XML metadata -XML per newspaper page	METS and ALTO schemas	~20TB	**FMP**

NLP Preprocessing





NLP Preprocessing



1. Divide the text in sentences

2. Each word of a sentence is tokenised

3. Each token is normalised

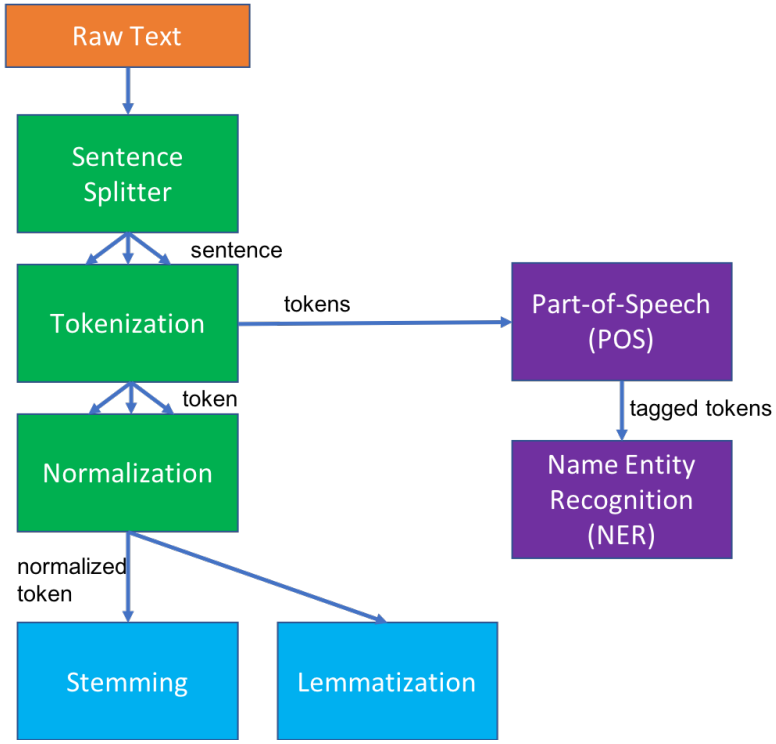
4.a) Stemming: reduces words to their roots

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)

NLP Preprocessing

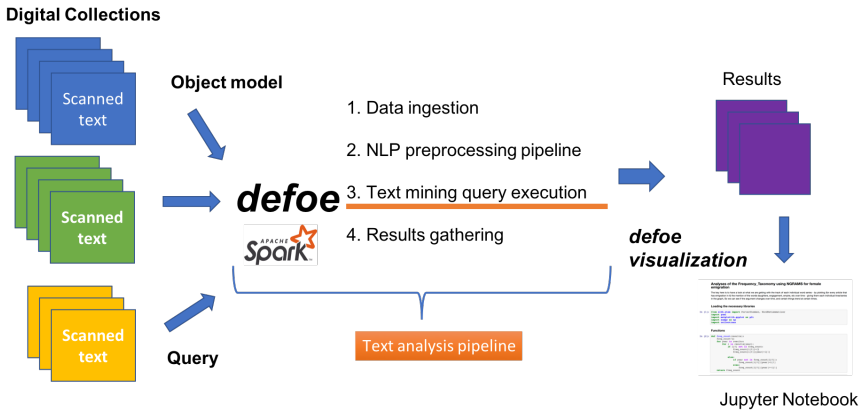


Introducing NLP to our queries: NLTK and spaCy

```
Sentence: "And devoted some time to social work in London."

|spaCy preprocessing|
 Word     Normaliz. Lemma     PoS      Tag    NER
 And      and       and       CCONJ    CC
 devoted  devoted   devote    VERB     VBN
 some     some      some      DET      DT
 time     time      time      NOUN     NN
 to       to        to        ADP      IN
 social   social    social    ADJ      JJ
 work     work      work      NOUN     NN
 in       in        in        ADP      IN
 London   london    London    PROPN    NNP    GPE
 .                  .         PUNCT    .
-------------------------------------------------------------
|NLTK preprocessing|
 Word     Normaliz. Lemma    Stem     PoSTag      NER
 And      and       and      and      CC       (S And/CC)
 devoted  devoted   devoted  devot    VBN      (S devoted/VBN)
 some     some      some     some     DT       (S some/DT)
 time     time      time     time     NN       (S (NP time/NN))
 to       to        to       to       TO       (S to/TO)
 social   social    social   social   JJ       (S social/JJ)
 work     work      work     work     NN       (S (NP work/NN))
 in       in        in       in       IN       (S in/IN)
 London   london    london   london   NNP   (S(GPE London/NNP))
.                  .        .          .      (S ./.)
```

Text Mining Queries



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

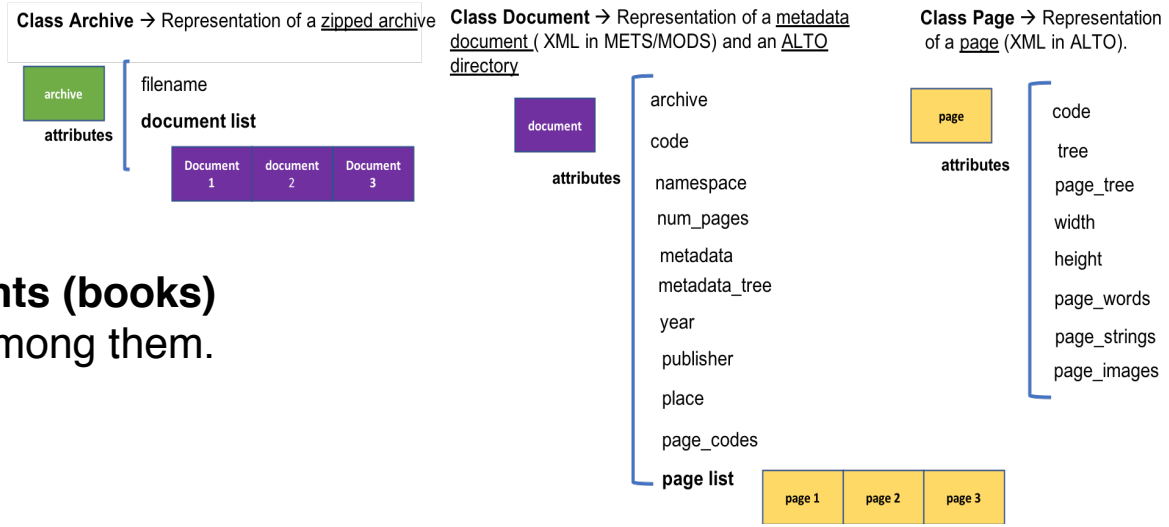
- keyword_by_year.py
- keyword_concordance_by_date.py
- normalize.py
- total_articles.py
- total_words.py

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.

```
# [archive, archive, ...]
documents = archives.flatMap(lambda archive: list(archive))
# [num_words, num_words, ...]
num_words = documents.map(lambda document: len(list(document.words())))
result = [documents.count(), num_words.reduce(add)]
return {"num_documents": result[0],
        "num_words": result[1]}
```

Sample results

Query over British Library Books

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

Installing defoe

Install Spark and Java 8

```
sudo apt install openjdk-8-jdk
wget http://apache.mirror.anlx.net/spark/spark-2.4.4/spark-2.4.4-bin-hadoop2.7.tgz
tar xvf spark-2.4.2-bin-hadoop2.7.tgz
```

Conda environment (Python 2.7)

```
git clone https://github.com/alan-turing-institute/defoe.git

conda create -n LwMpy27 python=2.7 anaconda

conda activate LwMpy27

conda update -n base -c defaults conda
```

Install dependencies

```
./requirements.sh

pip install Pillow==4.0.0
```

Install NLTK

```
    python

        import nltk

        nltk.download('wordnet')
```

https://github.com/alan-turing-institute/defoe/blob/master/docs/setup-local.md

Running defoe

Submit the source code to Spark along with information about your query:

```
spark-submit --py-files defoe.zip defoe/run_query.py <DATA_FILE> <MODEL_NAME> <QUERY_NAME> <QUERY_CONFIG
```

1. Data file: URLs or file paths

0000164- The Courier and Argus
0000187- The Bath Chronicle

Model Name: text model is to be used

papers

Query name: name of the python query module

defoe.papers.queries.articles_containing_words

Query configuration: (optional) - query-specific configuration file

queries/emigration.yml

Complete information:

https://github.com/alan-turing-institute/defoe/blob/master/docs/run-queries.md

Case Studies -Stranger Danger

Detect the origin of 'Stranger Danger' expression

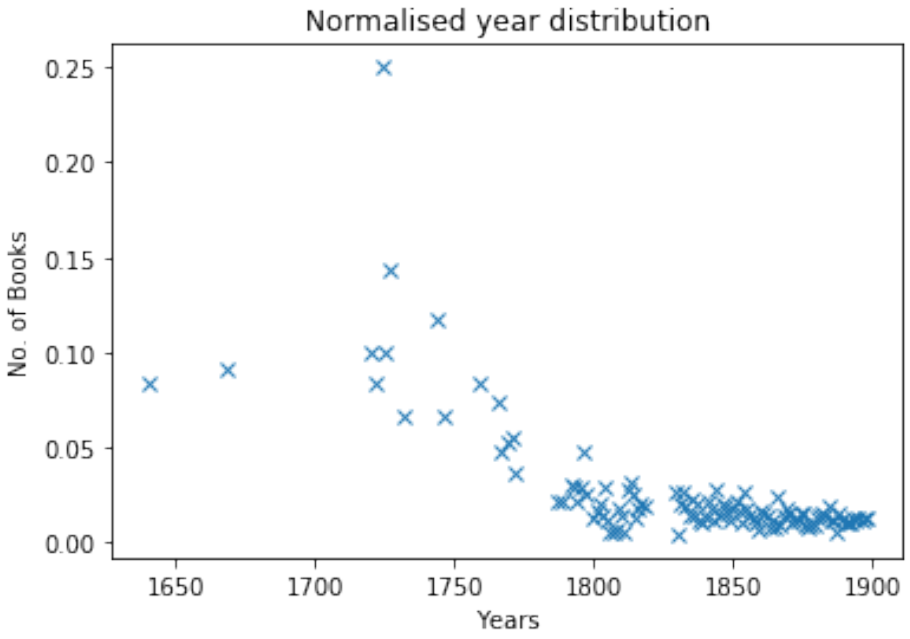
- ***colocates_by_year*** query:
  Searches for for sentences where the words "*stranger*" and "*danger*" (**matching criteria**) appear within the same sentence.
  Results are grouped by the publication dates.

- ***Normalise*** query:
  Counts total number of documents, pages and words per year

- Jupyter Notebook (*)
  - Compare results - plot them by year
  - Normalise the results
  - Sentiment analyses
  - Visualise which words appear more often near the phrase

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

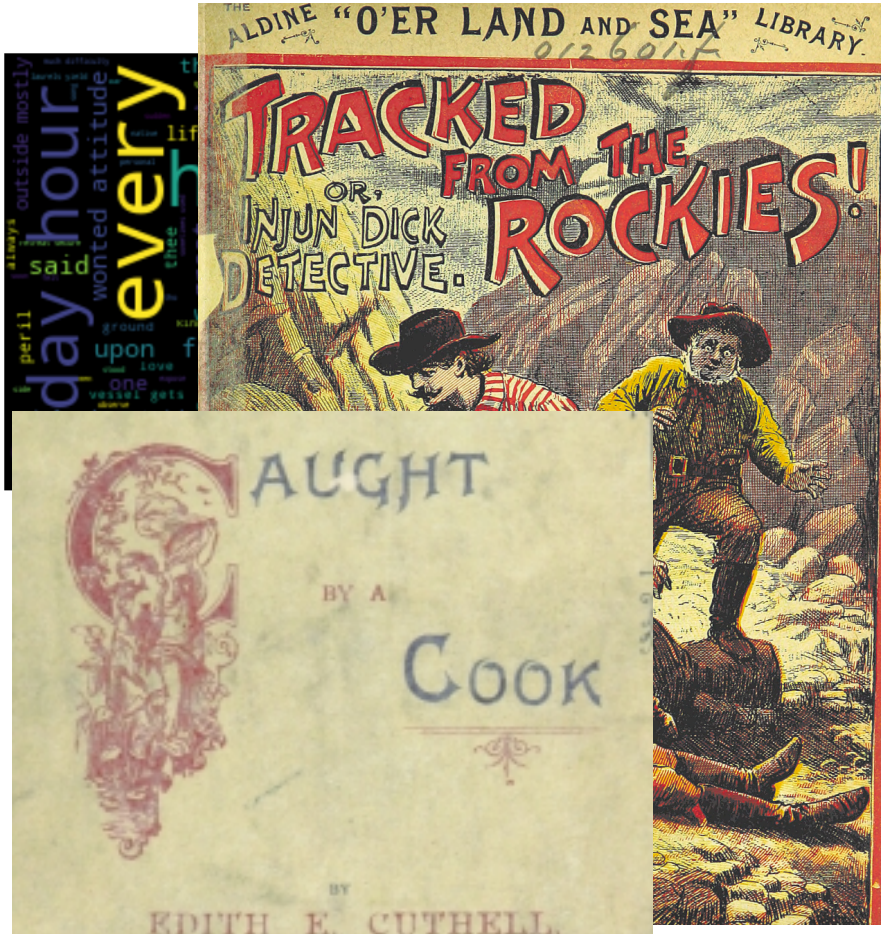
Case Studies -Stranger Danger



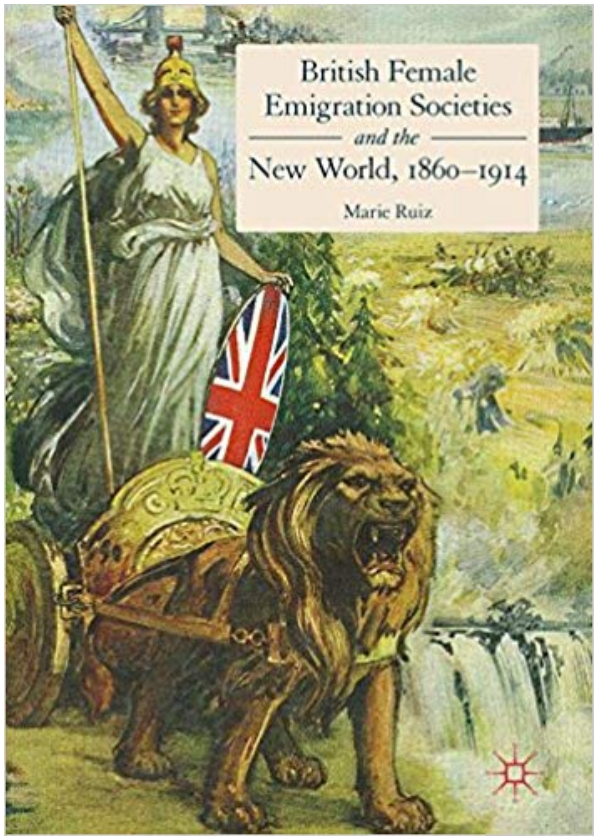
Visualization of how the 'Stranger Danger' terms are affected by the way that the number of books were published

Getting the first book, which both terms together:
"Caught by a Cook (1895)", Page: 105.
Sentence: "*who knew most of the villagers by sight perceived a stranger danger*."

Getting the total number of matching sentences, and the book which has the max. num of them:
Total sentences found is 1038
The book "*Aldine "O'er Land and Sea"* (1890) ,
has the max number of sentences: 29



Case Studies – Female Emigration



Mine the TDA and BLN archives for attitudes towards female emigration from Great Britain to the 'Colonies' and North America from 1850 to 1914

Normalised frequencies of female emigration societies:
keysentence_by_year & ***normalize*** queries

Normalised frequencies of taxonomy terms relating to female emigration:
`target_and_keywords_count_by_year & normalize` queries

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

Case Studies – Female Emigration

Taxonomy terms

emigration
bookkeeping
Colony
Colonies
Colonial
daughters
engagement
empire
British empire
failure
female labourer
female welfare
feminine
genteel
good character
governess
guardian
guardianship
happiness
hardship

Female taxonomy terms co-located with "emigration"

1850:
- [teacher, 23]
- [loan, 138]
- [mother, 102]
- [happiness, 15]
- [genteel, 6]
- [marriage, 57]
- [respectable, 177]

1851:
- [indecent, 2]
- [superintendence, 13]
- [loan, 130]
- [suitable, 54]
- [colony, 768]
- [bookkeeping, 1]
- [success, 78]

Using TDA corpus:
Only newspapers
From 1850 to 1914.

Frequencies of female societies

1850:
- [governess benevolent institution, 1]
- [emigration agent, 17]
- [emigration scheme, 4]
- [colonial land and emigration commission, 2]
- [family colonization loan society, 2]

Societies

South African Colonisation Society
Juvenile Emigration Society
Child Emigration Society
Church Emigration Society
Self Help Emigration Society
East End Emigration Fund
Church Army Emigration Department
Carlton Emigration Society to Canada

Total number of issues, articles and words per year

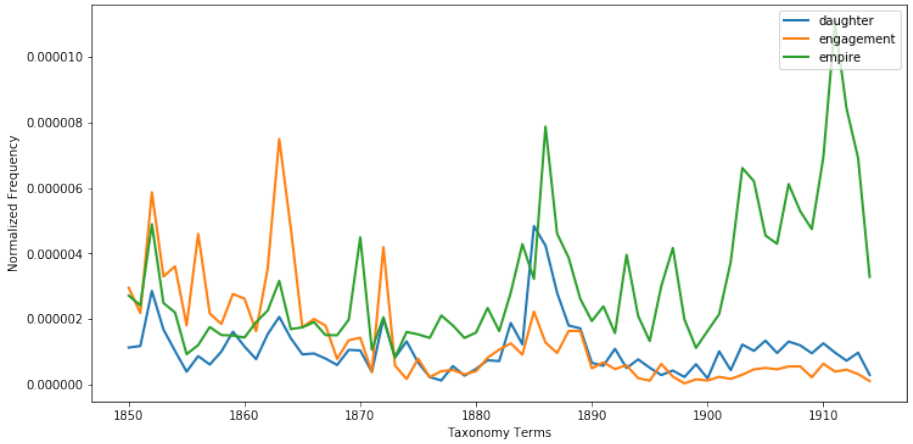
1850: [313, 22288, 33449528]
1851: [313, 21247, 33843690]
1852: [314, 19586, 33903669]

Case Studies – Female Emigration

Taxonomy terms

emigration
bookkeeping
Colony
Colonies
Colonial
daughters
engagement
empire
British empire
failure
female labourer
female welfare
feminine
genteel
good character
governess
guardian
guardianship
happiness
hardship

Normalised N-grams of female taxonomy terms co-located with "emigration"



Using TDA corpus:
Only newspapers
From 1850 to 1914.

Normalised frequencies of female societies

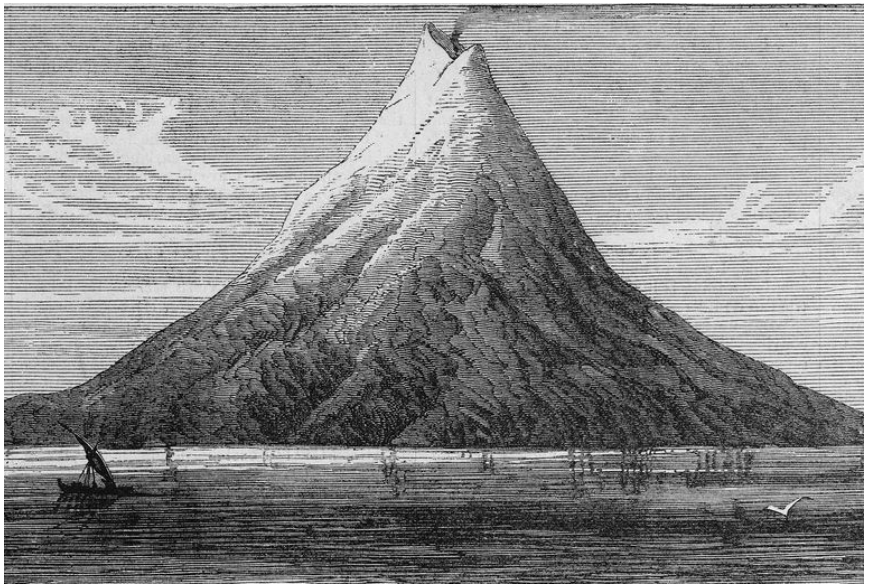
Societies

South African Colonisation Society
Juvenile Emigration Society
Child Emigration Society
Church Emigration Society
Self Help Emigration Society
East End Emigration Fund
Church Army Emigration Department
Carlton Emigration Society to Canada

Case Studies -Eruption of Krakatoa Volcano in 1883

Track copying and transmission of text

Krakatoa (Krakatau in Indonesian) erupted over 26-27th August 1883 and was one of the most spectacular volcanic eruptions in contemporary times.



Identify this eruption using
BLN, TDA and NZPP papers from late 1883

`keyword_and_concordance_by_date` `query:` searches for occurrences of "*krakatoa*" and "*krakatua*" and returns information on each matching article.

https://github.com/alan-turing-institute/defoe_visualization/tree/master/Krakatoa_1883

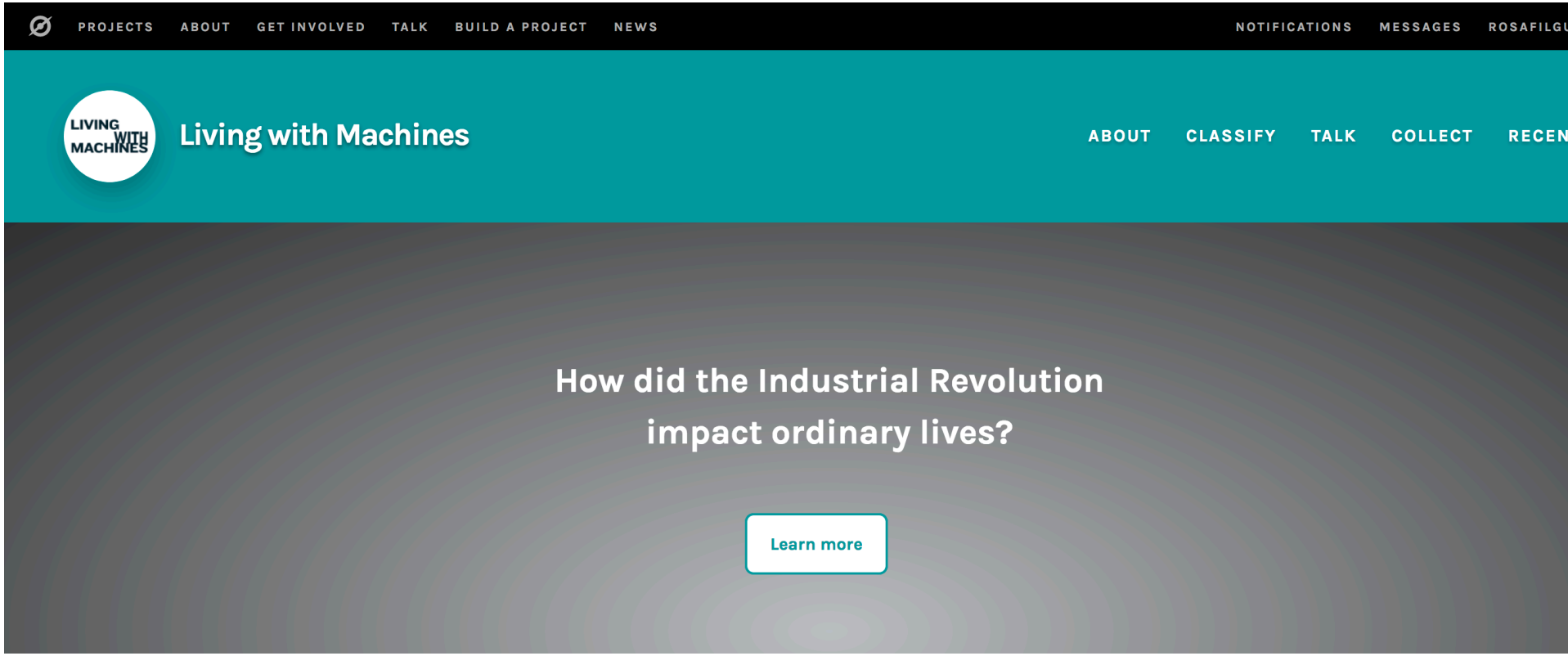
Case Studies -Eruption of Krakatoa Volcano in 1883

	date	newspaper title	search term	sequence	text
2	1883-05-24	0000453- The Evening Telegraph	krakatoa	1	VOLCANIC ERUPTION. Rata via, Thursday. A violent eruption ha* be
3	1883-05-25	0000327- The Derby Daily Telegraph	krakatoa	1	VOLCANIC ERUPTION IN THE STRAITS OF SUNDA. A Reuter's teleg
4	1883-08-27	0000321- Nottingham Evening Post	krakatoa	1	ALARMING VOLCANIC DISTURBANCES. DISTURBANCES. A Calami
5	1883-08-27	0000325- The Citizen Gloucester	krakatoa	1	TERRIBLE VOLCANIC DISTURBANCE IN THE WEST INDIES : A VILL
6	1883-08-28	0000321- Nottingham Evening Post	krakatoa	1	THE VOLCANIC ERUPTION. r • -Manager. 2473 c " OFFICII,"^s"~ >
7	1883-08-28	0000452- Edinburgh Evening News	krakatoa	1	GREAT VOLCANIC ERUPTION IN THE EAST INDIES. A Keuter's teleg
8	1883-08-28	0000325- The Citizen Gloucester	krakatoa	1	THE VOLCANIC DISTURBANCES IN THE EAST INDIES. Batavia, Aug
9	1883-08-28	0000327- The Derby Daily Telegraph	krakatoa	1	SUMMARY. ®be Her bp ®axlg ®eUjraplj DERBY, August 28, 1883. Tb
10	1883-08-29	0000327- The Derby Daily Telegraph	krakatoa	1	LATEST NEWS. "Telegraph" Officii, 1.0 P.M. MINISTERIAL CRISIS IN
11	1883-08-29	0000325- The Citizen Gloucester	krakatoa	1	THE VOLCANIC DISTURBANCES IN JAVA: A TOWN DESTROYED BY
12	1883-08-29	0000321- Nottingham Evening Post	krakatoa	1	A TOWN COMPLETELY DESTROYED. S^OIALjEDITION.! POST" OFF

~50 newspaper articles from all over England
Rich data set where we can track copying and transmission of text.

Case Studies - Zooniverse and *Industrial accidents*

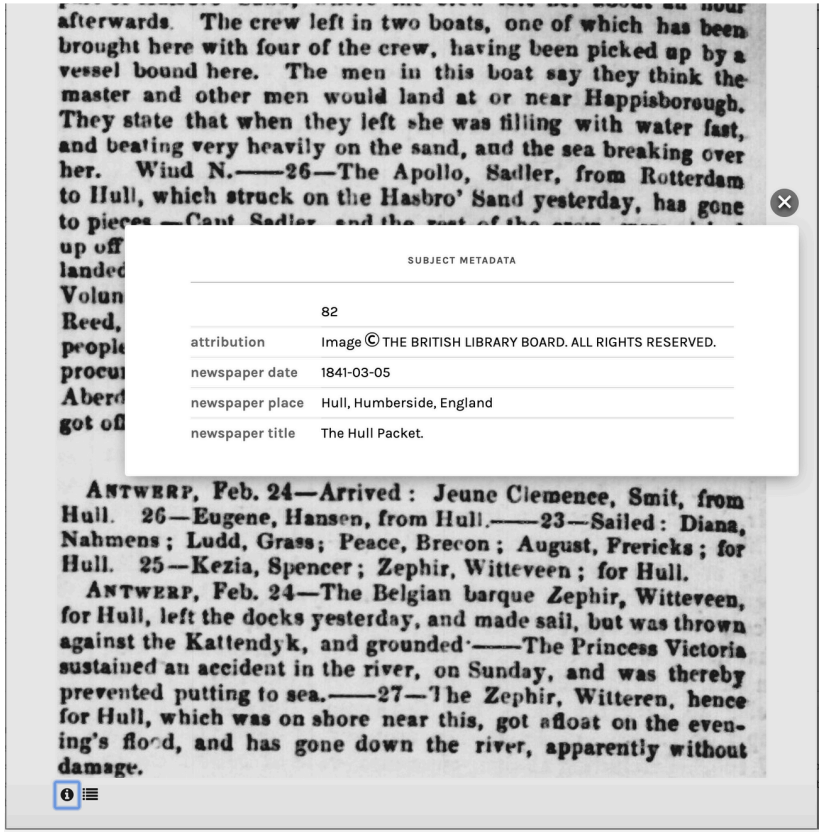
Zooniverse- Largest and most popular citizen science projects

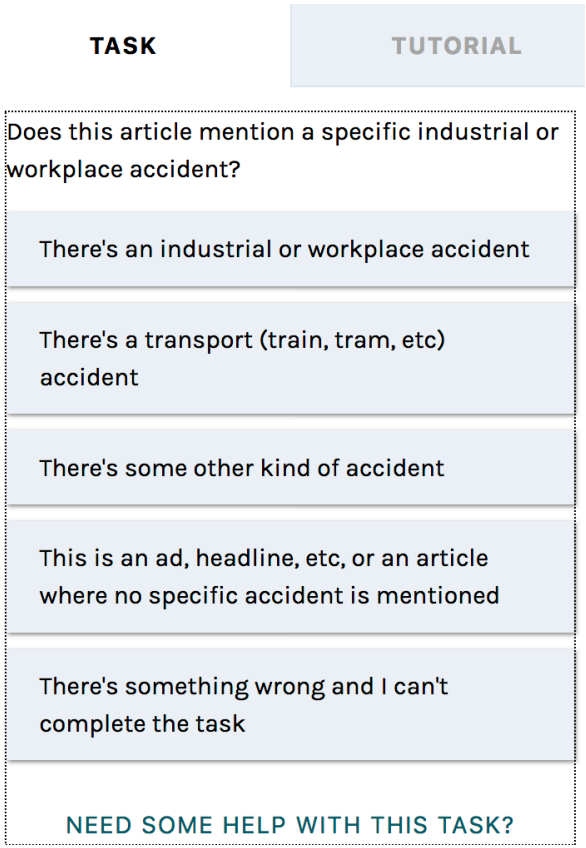


Victorian newspapers (FMP newspapers) didn't talk about 'industrial accidents', but they are full of accounts of workplace injuries linked to machines.

https://www.zooniverse.org/projects/bldigital/living-with-machines

Case Studies - Zooniverse and *Industrial accidents*





Victorian newspapers (FMP newspapers) didn't talk about 'industrial accidents', but they are full of accounts of workplace injuries linked to machines.

https://www.zooniverse.org/projects/bldigital/living-with-machines

Case Studies - Zooniverse and *Industrial accidents*

Defoe – Filtering and Cropping article's images – FMP newspapers

Target_segmentation : Crops articles textblocks that at least have either **MACHINE** or **MACHINERY** (**target words**), **AND** any of the **keywords** (*accident, crush, smash, etc.*) listed in the lexicon.

```
spark-submit --py-files defoe.zip defoe/run_query.py data.txt
fmp defoe.fmp.queries.target_segmentation queries/accident.yml
```

Results: Metadata file (article_id, publication, year)
and **Images files (upload to Zooniverse)**

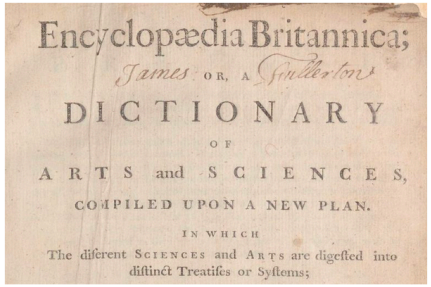
National Library Scottish (**New!**)

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

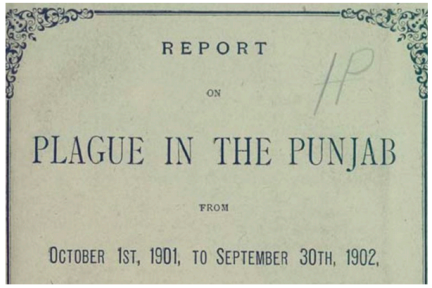
HOME ABOUT DATA TOOLS CONTACT

Digitised collections

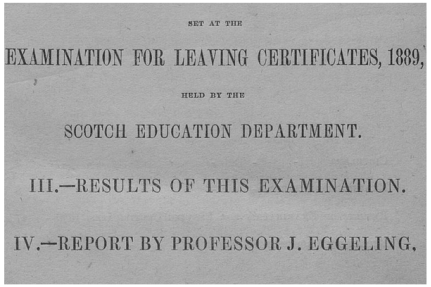
Download the ALTO, METS, image and plain text files for our digitised collections.



Encyclopaedia Britannica, 1768-1860



A Medical History of British India



Scottish School Exam Papers, 1888-1963

National Library Scottish (**New!**)

Original OCR: no clean-up

155,388 ALTO XML files at page level

155,388 image files

METS metadata files at item level

19,257,785 lines and 166,729,009 words

Covers years 1768-1860

https://github.com/alan-turing-institute/defoe/tree/master/defoe/nls

Physical representation: one XML METS per volume and a XML ALTO per page

Computing Environments – defoe - portable

Cray Urika-GX system:

- High-performance analytics cluster
- **Apache Spark**, Apache Hadoop, Jupyter Notebooks, etc.
- 12 computing nodes: 36 CPUs and 256GB
- 60TB of storage – HDFS & Lustre

EDDIE :

- University of Edinburgh HPC cluster.
- 7000 Intel cores with up to 3TB per compute node

Microsoft Azure

- Alan Turing Allocation
- HDInsight: Cloud-based service for big data analytics -- it includes Spark
- Apache Livy: Submitting (Spark) Jobs from Anywhere

Conclusions

- New digital toolbox for extracting knowledge from historical data.
- Enables running text analyses across large collections in parallel.
- Offers a rich set of text mining queries.
- Includes NLP prepossessing techniques to mitigate OCR errors.
- Tested portability on different computing environments and digital collections.

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