Text Analysis with NLTK

Week 1

6 November 2020

CDCS Python Course Series

with Lucy Havens

NLTK: A Quick Recap

Natural Language Toolkit

Examples of common sources of natural language (unstructured data):

- Websites
- News articles
- Books
- Audio transcriptions
- Social media posts

NLTK

Built-in methods for getting familiar with a text:

```
.concordance(“word”, lines=20)
.similar(“word”)
.common_contexts([“list”, “of”, “words”])
.dispersion_plot([“list”, “of”, “words”])
```

Reference: https://www.nltk.org/book/ch01.html

NLTK

Built-in functions and methods for summarizing a text:

```
len(text)
sorted(vocab_of_text)
.count(“word”)
```

Reference: https://www.nltk.org/book/ch01.html

Assignment

Prework for those who haven’t used Python or Jupyter Notebooks before

Helpful Resources - **for reference**, not required reading

Steps 1-5: Independent learning
Step 6: A tutorial to complete in your own Jupyter Notebook

Go Further - **optional** reading assignment

How did it go?

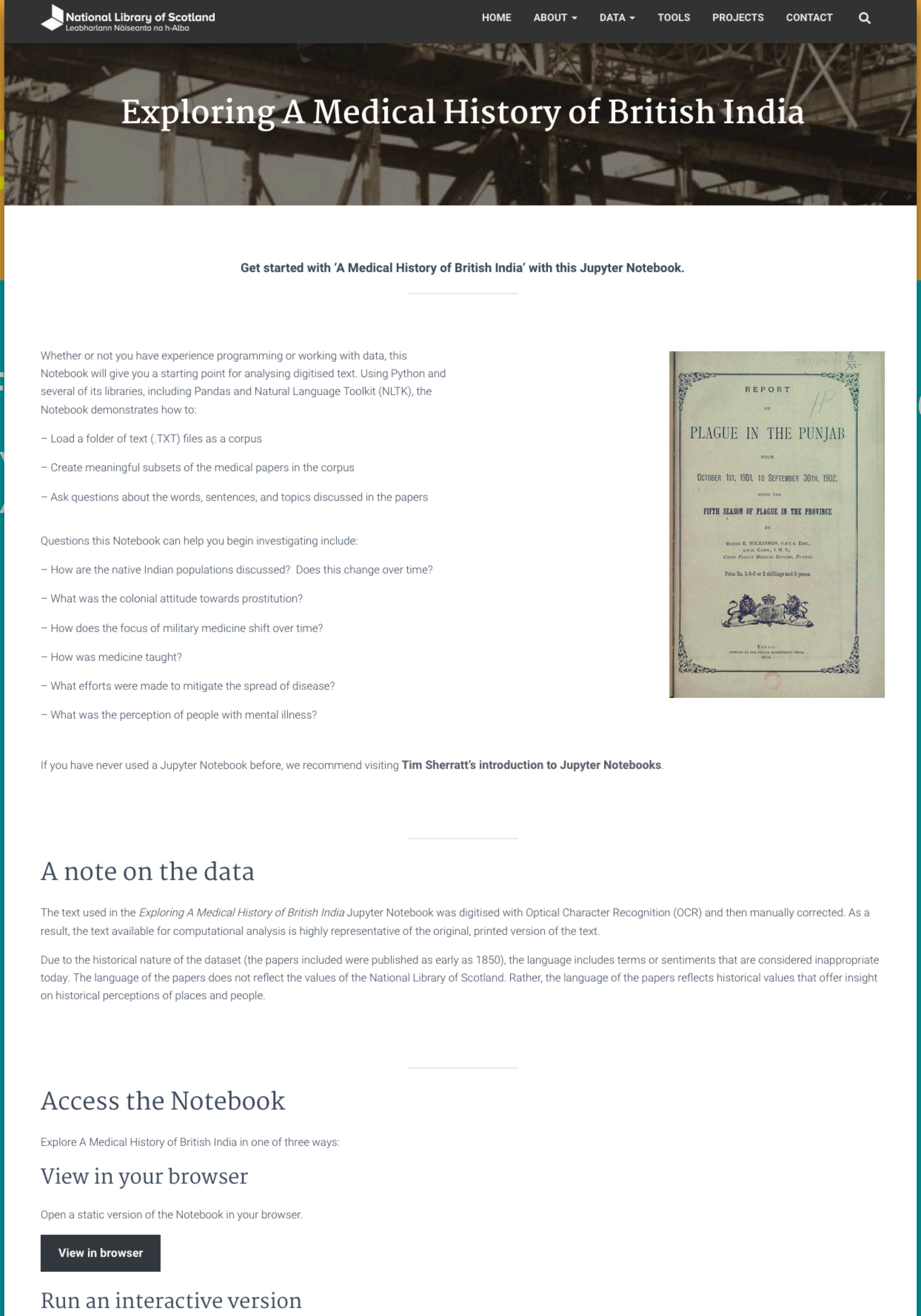
DEMO

Assignment

What was the name of the dataset used in *Library Carpentry: Text & Data Mining* (assigned in #6)?

Assignment

What was the name of … *entry: Text & Data Mining* (assigned in #6)



Exploring A Medical History of British India

Get started with 'A Medical History of British India' with this Jupyter Notebook.

Whether or not you have experience programming or working with data, this Notebook will give you a starting point for analysing digitised text. Using Python and several of its libraries, including Pandas and Natural Language Toolkit (NLTK), the Notebook demonstrates how to:

– Load a folder of text (.TXT) files as a corpus

– Create meaningful subsets of the medical papers in the corpus

– Ask questions about the words, sentences, and topics discussed in the papers

Questions this Notebook can help you begin investigating include:

– How are the native Indian populations discussed? Does this change over time?

– What was the colonial attitude towards prostitution?

– How does the focus of military medicine shift over time?

– How was medicine taught?

– What efforts were made to mitigate the spread of disease?

– What was the perception of people with mental illness?

REPORT
ON
PLAGUE IN THE PUNJAB
FROM
OCTOBER 1ST, 1901, TO SEPTEMBER 30TH, 1902,
BEING THE
FIFTH SEASON OF PLAGUE IN THE PROVINCE
BY
MAJOR E. WILKINSON, F.R.C.S. ENG.,
D.P.H. CAMB., I. M. S.,
CHIEF PLAGUE MEDICAL OFFICER, PUNJAB.
Price Rs. 1-9-0 or 2 shillings and 6 pence.
Lahore:
PRINTED AT THE PUNJAB GOVERNMENT PRESS,
1904.

If you have never used a Jupyter Notebook before, we recommend visiting **Tim Sherratt's introduction to Jupyter Notebooks**.

A note on the data

The text used in the *Exploring A Medical History of British India* Jupyter Notebook was digitised with Optical Character Recognition (OCR) and then manually corrected. As a result, the text available for computational analysis is highly representative of the original, printed version of the text.

Due to the historical nature of the dataset (the papers included were published as early as 1850), the language includes terms or sentiments that are considered inappropriate today. The language of the papers does not reflect the values of the National Library of Scotland. Rather, the language of the papers reflects historical values that offer insight on historical perceptions of places and people.

Access the Notebook

Explore A Medical History of British India in one of three ways:

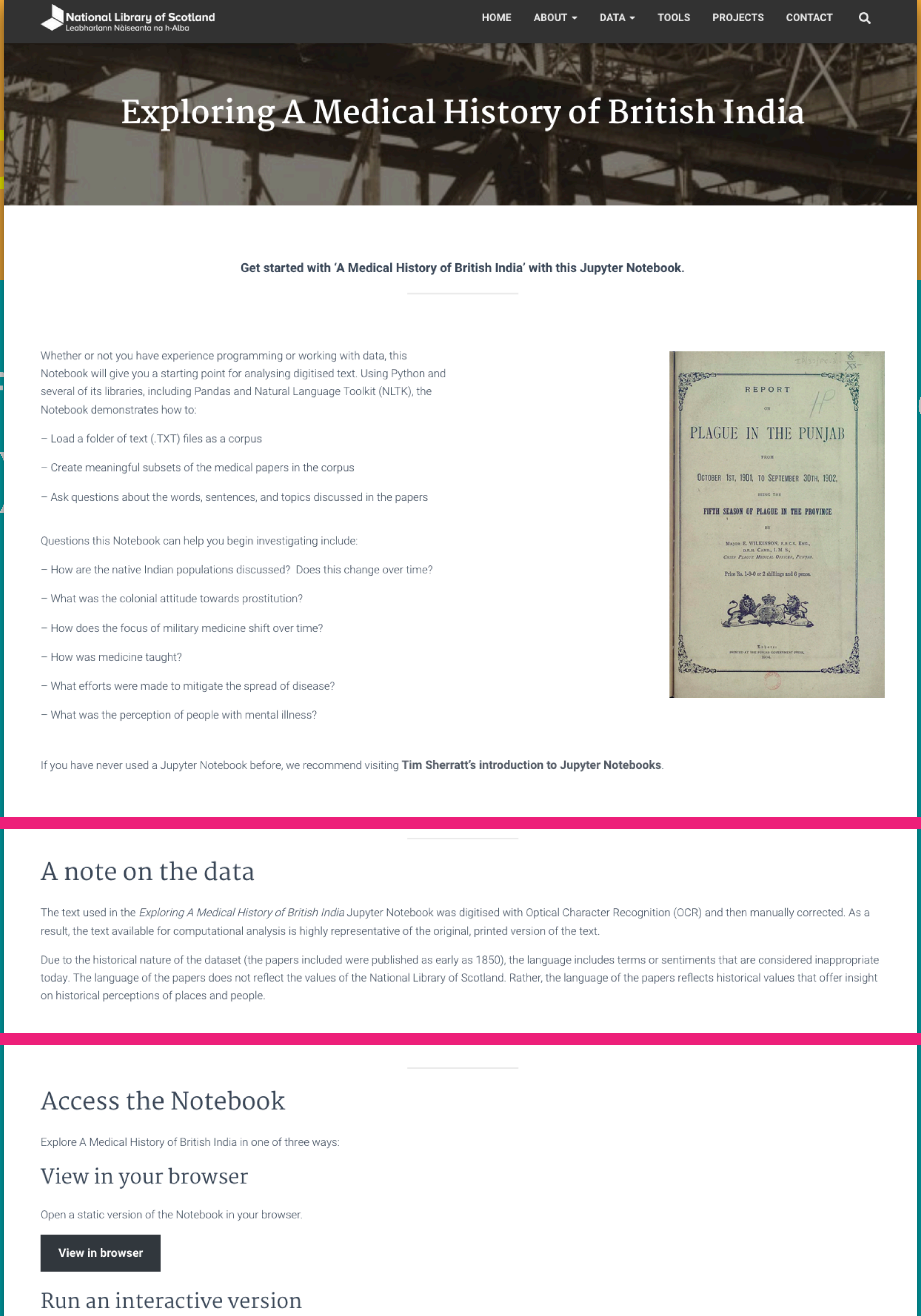
View in your browser

Open a static version of the Notebook in your browser.

Run an interactive version

Assignment

What was the name of *entry: Text & Data Mining* (assigned in #6)


National Library of Scotland
Leabharlann Nàiseanta na h-Alba
HOME
ABOUT
DATA
TOOLS
PROJECTS
CONTACT
REPORT ON PLAGUE IN THE PUNJAB
FROM OCTOBER 1st, 1901, TO SEPTEMBER 30th, 1902,
BEING THE FIFTH SEASON OF PLAGUE IN THE PROVINCE

Exploring A Medical History of British India

Get started with 'A Medical History of British India' with this Jupyter Notebook.

Whether or not you have experience programming or working with data, this Notebook will give you a starting point for analysing digitised text. Using Python and several of its libraries, including Pandas and Natural Language Toolkit (NLTK), the Notebook demonstrates how to:

– Load a folder of text (.TXT) files as a corpus

– Create meaningful subsets of the medical papers in the corpus

– Ask questions about the words, sentences, and topics discussed in the papers

Questions this Notebook can help you begin investigating include:

– How are the native Indian populations discussed? Does this change over time?

– What was the colonial attitude towards prostitution?

– How does the focus of military medicine shift over time?

– How was medicine taught?

– What efforts were made to mitigate the spread of disease?

– What was the perception of people with mental illness?

If you have never used a Jupyter Notebook before, we recommend visiting **Tim Sherratt's introduction to Jupyter Notebooks**.

A note on the data

The text used in the *Exploring A Medical History of British India* Jupyter Notebook was digitised with Optical Character Recognition (OCR) and then manually corrected. As a result, the text available for computational analysis is highly representative of the original, printed version of the text.

Due to the historical nature of the dataset (the papers included were published as early as 1850), the language includes terms or sentiments that are considered inappropriate today. The language of the papers does not reflect the values of the National Library of Scotland. Rather, the language of the papers reflects historical values that offer insight on historical perceptions of places and people.

Access the Notebook

Explore A Medical History of British India in one of three ways:

View in your browser

Open a static version of the Notebook in your browser.

View in browser

Run an interactive version

Assignment

Exploring A Medical History of British India

Get started with 'A Medical History of British India' with this Jupyter Notebook.

A note on the data

The text used in the *Exploring A Medical History of British India* Jupyter Notebook was digitised with Optical Character Recognition (OCR) and then manually corrected. As a result, the text available for computational analysis is highly representative of the original, printed version of the text.

Due to the historical nature of the dataset (the papers included were published as early as 1850), the language includes terms or sentiments that are considered inappropriate today. The language of the papers does not reflect the values of the National Library of Scotland. Rather, the language of the papers reflects historical values that offer insight on historical perceptions of places and people.

Access the Notebook

Explore A Medical History of British India in one of three ways:

View in your browser

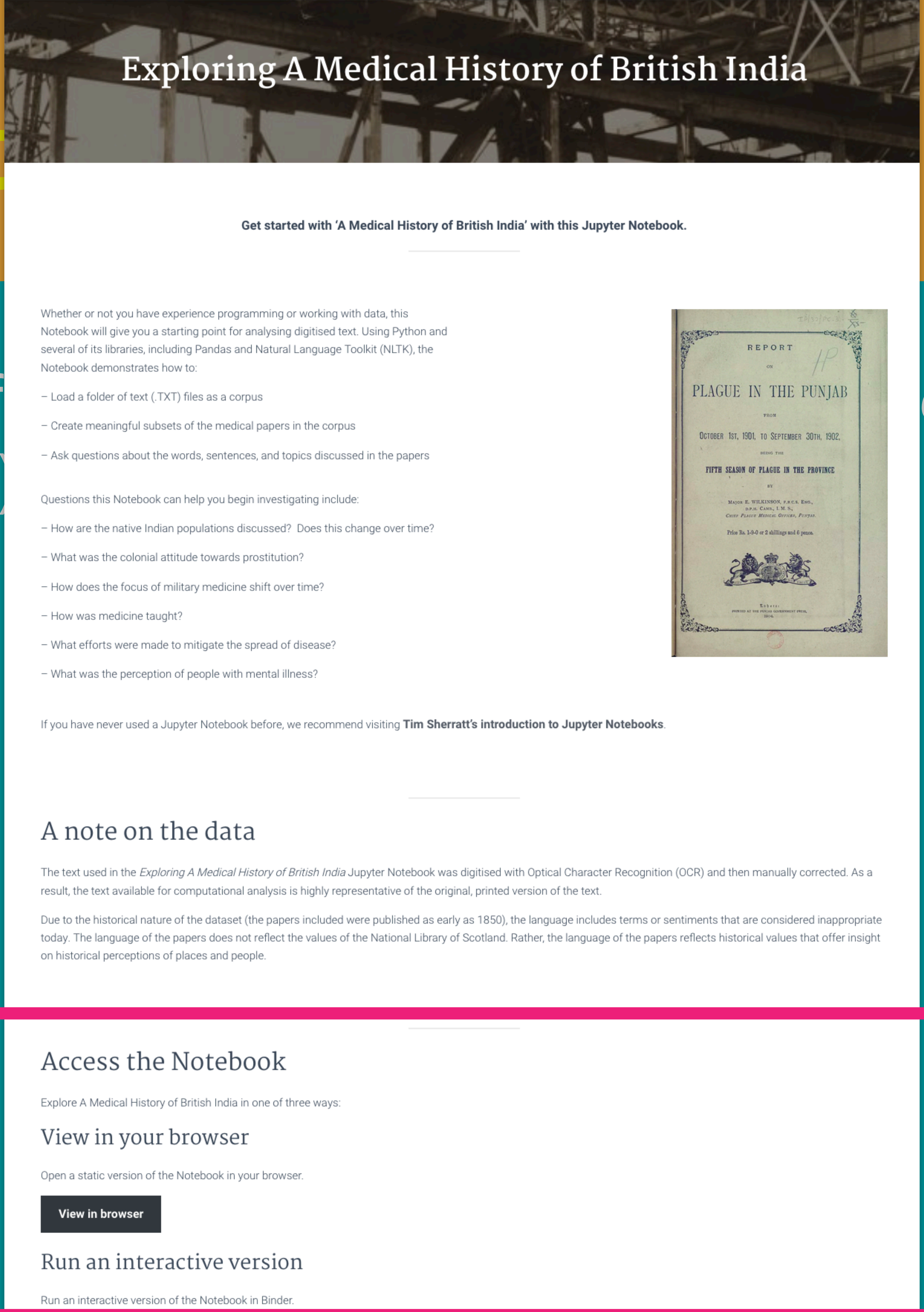
Open a static version of the Notebook in your browser.

View in browser

Run an interactive version

Assignment

What was the name of *entry: Text & Data Mining* (assigned in #6)



Exploring A Medical History of British India

Get started with 'A Medical History of British India' with this Jupyter Notebook.

Whether or not you have experience programming or working with data, this Notebook will give you a starting point for analysing digitised text. Using Python and several of its libraries, including Pandas and Natural Language Toolkit (NLTK), the Notebook demonstrates how to:

– Load a folder of text (.TXT) files as a corpus

– Create meaningful subsets of the medical papers in the corpus

– Ask questions about the words, sentences, and topics discussed in the papers

Questions this Notebook can help you begin investigating include:

– How are the native Indian populations discussed? Does this change over time?

– What was the colonial attitude towards prostitution?

– How does the focus of military medicine shift over time?

– How was medicine taught?

– What efforts were made to mitigate the spread of disease?

– What was the perception of people with mental illness?

REPORT
ON
PLAGUE IN THE PUNJAB
FROM
October 1st, 1901, to September 30th, 1902,
BEING THE
FIFTH SEASON OF PLAGUE IN THE PROVINCE
BY
Major E. Wilkinson, F.R.C.S. Eng.,
D.P.H. Camb., I. M. S.,
Chief Plague Medical Officer, Punjab.
Price Rs. 1-9-0 or 2 shillings and 6 pence.
Lahore:
PRINTED AT THE PUNJAB GOVERNMENT PRESS,
1904.

If you have never used a Jupyter Notebook before, we recommend visiting **Tim Sherratt's introduction to Jupyter Notebooks**.

A note on the data

The text used in the *Exploring A Medical History of British India* Jupyter Notebook was digitised with Optical Character Recognition (OCR) and then manually corrected. As a result, the text available for computational analysis is highly representative of the original, printed version of the text.

Due to the historical nature of the dataset (the papers included were published as early as 1850), the language includes terms or sentiments that are considered inappropriate today. The language of the papers does not reflect the values of the National Library of Scotland. Rather, the language of the papers reflects historical values that offer insight on historical perceptions of places and people.

Access the Notebook

Explore A Medical History of British India in one of three ways:

View in your browser

Open a static version of the Notebook in your browser.

Run an interactive version

Run an interactive version of the Notebook in Binder.

Up Next

use NLTK to analyze a text of your choice

Thanks everyone!

Next course meeting: Monday, 10:00-11:00 AM BST

Office hours available on Wednesday (30 minutes)

To schedule, please message me on Teams!