Kraken Exploration with English Text

Yian Wang

October 17, 2020

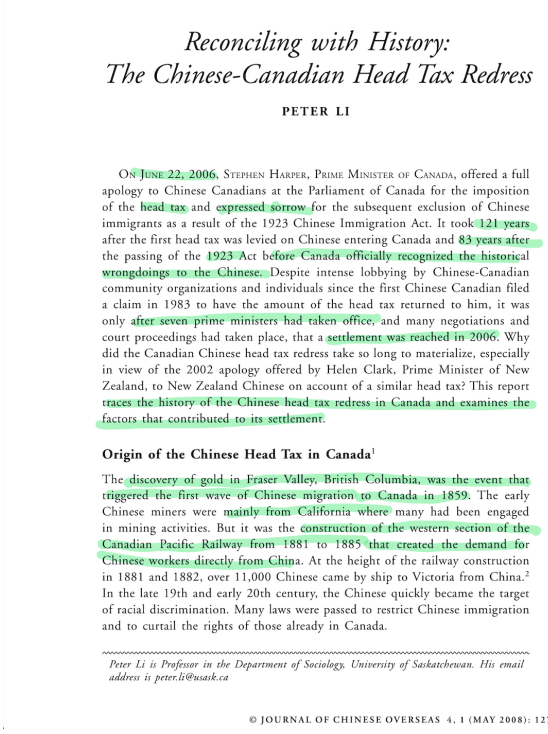
1. Introduction

Kraken is an easy-to-follow system used for Optical Character Recognition (OCR) [1]. The purpose of this exploration is to look at a one-page, one-column academic sample text to ensure kraken works sufficiently for English text before training it for Japanese characters.

2. Data

2.1 Sample Text

The sample "Reconciling with History: The Chinese-Canadian Head Tax Redress" by Peter Li [2] contains varying types of text - the title, author, main text, subheading, and footnote, and watermark are all different. Thus, an area of interest will be to see how kraken interprets the different types of text, in particular, the 1 superscript in the subheading "Origin of the Chinese Head Tax in Canada".
Keeping the green highlighted areas in the document tests the effectiveness of binarization in kraken.
As per the tutorial, the sample is imported as a tif filetype.



Reconciling with History:
The Chinese-Canadian Head Tax Redress

PETER LI

On June 22, 2006, Stephen Harper, Prime Minister of Canada, offered a full apology to Chinese Canadians at the Parliament of Canada for the imposition of the head tax and expressed sorrow for the subsequent exclusion of Chinese immigrants as a result of the 1923 Chinese Immigration Act. It took 121 years after the first head tax was levied on Chinese entering Canada and 83 years after the passing of the 1923 Act before Canada officially recognized the historical wrongdoings to the Chinese. Despite intense lobbying by Chinese-Canadian community organizations and individuals since the first Chinese Canadian filed a claim in 1983 to have the amount of the head tax returned to him, it was only after seven prime ministers had taken office, and many negotiations and court proceedings had taken place, that a settlement was reached in 2006. Why did the Canadian Chinese head tax redress take so long to materialize, especially in view of the 2002 apology offered by Helen Clark, Prime Minister of New Zealand, to New Zealand Chinese on account of a similar head tax? This report traces the history of the Chinese head tax redress in Canada and examines the factors that contributed to its settlement.

Origin of the Chinese Head Tax in Canada[1]

The discovery of gold in Fraser Valley, British Columbia, was the event that triggered the first wave of Chinese migration to Canada in 1859. The early Chinese miners were mainly from California where many had been engaged in mining activities. But it was the construction of the western section of the Canadian Pacific Railway from 1881 to 1885 that created the demand for Chinese workers directly from China. At the height of the railway construction in 1881 and 1882, over 11,000 Chinese came by ship to Victoria from China.[2] In the late 19th and early 20th century, the Chinese quickly became the target of racial discrimination. Many laws were passed to restrict Chinese immigration and to curtail the rights of those already in Canada.

Peter Li is Professor in the Department of Sociology, University of Saskatchewan. His email address is peter.li@usask.ca

© JOURNAL OF CHINESE OVERSEAS 4, 1 (MAY 2008): 127–140

Figure 1. Sample Input Text: Reconciling with History

2.2 en_best.mlmodel

In the current kraken version to date, this model was not included but was required to run the code. It must be downloaded and imported separately (**click here to download** - see mittagessen's comment on Feb 15, 2019).

3. Kraken Exploration

3.1 Local

The kraken tutorial is easily followed locally; however, for this process to be as accessible and reproducible as possible, this exploration focuses on getting the process to work on some cloud platform such as Jupyter Hub or Google Colab.

3.2 Jupyter Notebook/Jupyter Hub

JupyterLab on Jupyter Hub is a free cloud platform to share code.

3.2.1 Error

Attempting to use kraken in the JupyterLab terminal, the shell runs out of memory before the package fully installs. The first command

```
pip install kraken
```

reaches a certain point in memory usage (1.84/2.00 GB) and proceeds to quit (See Appendix Figure A). The conclusion is that Jupyter Hub does not have adequate memory space so other platforms need to be explored.

3.3 Google Colab

Google Golab was chosen next as it has more space than JupyterHub.

3.3.1 Pricing

While Google Colab is free for up to 12GB of memory and run time of 12 hours, paying $9.99 a month gives access to 25GB and up to 24 hours of run time [3]. Though the free version is sufficient for the scope of this exploration task, this removes any fear of running out of space on this platform for the time being. Nonetheless, the discussion of pricing is important in order to gauge whether it's worth it to use the platform if/when there is a high volume of data.

3.3.2 Method and Outputs

First install kraken:

```
pip install kraken
```

Before beginning the process, upload the tif file of the desired text, as well as the `en_best.mlmodel` file to the Colab environment.

Now, binarize the image to remove any color to allow kraken to analyze it:

```
!kraken -i image.tif bw.png binarize
```

Then, take the binarized image output and turn it into a txt file:

```
!kraken -i bw.png image.txt binarize segment ocr -m en_best.mlmodel
```

Alternatively, it is possible to binarize and convert to a txt file in one line:

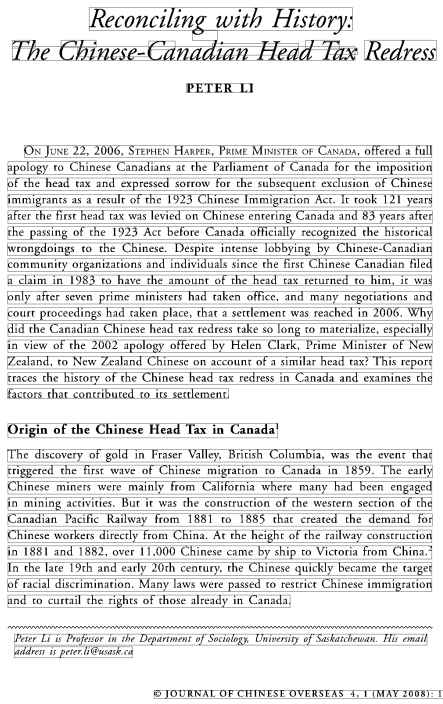
```
!kraken -i image.tif onestep_image.txt binarize segment ocr -m en_best.mlmodel
```

The resulting txt file is promising, with minimal errors, mostly in the title (See Appendix Figure B).
To analyse the issue, check the line boundaries to see if kraken detected the title area (and other areas of error) correctly. A json file of the boundaries is obtained and plotted around the text to see if there exist any unexpected boundaries.

Get the json file as follows:

```
!kraken -i bw.png lines.json segment
```

Then run a function `show_boxes` (see Appendix 6.1) to plot these boundaries. The output is as follows:



Reconciling with History:
The Chinese-Canadian Head Tax Redress

PETER LI

On June 22, 2006, Stephen Harper, Prime Minister of Canada, offered a full apology to Chinese Canadians at the Parliament of Canada for the imposition of the head tax and expressed sorrow for the subsequent exclusion of Chinese immigrants as a result of the 1923 Chinese Immigration Act. It took 121 years after the first head tax was levied on Chinese entering Canada and 83 years after the passing of the 1923 Act before Canada officially recognized the historical wrongdoings to the Chinese. Despite intense lobbying by Chinese-Canadian community organizations and individuals since the first Chinese Canadian filed a claim in 1983 to have the amount of the head tax returned to him, it was only after seven prime ministers had taken office, and many negotiations and court proceedings had taken place, that a settlement was reached in 2006. Why did the Canadian Chinese head tax redress take so long to materialize, especially in view of the 2002 apology offered by Helen Clark, Prime Minister of New Zealand, to New Zealand Chinese on account of a similar head tax? This report traces the history of the Chinese head tax redress in Canada and examines the factors that contributed to its settlement.

Origin of the Chinese Head Tax in Canada[1]

The discovery of gold in Fraser Valley, British Columbia, was the event that triggered the first wave of Chinese migration to Canada in 1859. The early Chinese miners were mainly from California where many had been engaged in mining activities. But it was the construction of the western section of the Canadian Pacific Railway from 1881 to 1885 that created the demand for Chinese workers directly from China. At the height of the railway construction in 1881 and 1882, over 11,000 Chinese came by ship to Victoria from China.[2] In the late 19th and early 20th century, the Chinese quickly became the target of racial discrimination. Many laws were passed to restrict Chinese immigration and to curtail the rights of those already in Canada.

Peter Li is Professor in the Department of Sociology, University of Saskatchewan. His email address is peter.li@usask.ca

© JOURNAL OF CHINESE OVERSEAS 4, 1 (MAY 2008): 127–140

Figure 2. Sample Text with Line Boundaries

It can be see that kraken evaluated the boundaries well in this sample, with exception to the title, which has many strange boundary boxes around it. This explains the discrepancy in the produced txt file.

4. Evaluation

Kraken was fairly easy-to-use, and the online tutorial was very intuitive. It identified the words in the sample correctly with only minor errors.
However, the boundaries were found at a line level, rather than a word or character level, which is less effective and sensitive. Additionally, the `en_best.mlmodel` file was not installed with kraken in the command line, though that issue was easily solved.

5. Sources

1. kraken — kraken 2.0.5-4-gbb42ba5 documentation [Internet]. Kraken.re. 2015 [cited 2020 Oct 18]. Available from: http://kraken.re/index.html

2. Li P. Reconciling with History: The Chinese-Canadian Head Tax Redress. Journal of Chinese Overseas [Internet]. 2008 [cited 2020 Oct 18];4(1):127–40. Available from: https://brill.com/view/journals/jco/4/1/article-p127_9.xml DOI: 10.1163/179325408788691507

3. Google newly launches Colab Pro! - comparison of Colab and Colab pro · Buomsoo Kim [Internet]. Github.io. 2020 [cited 2020 Oct 18]. Available from: https://buomsoo-kim.github.io/colab/2020/03/15/Google-newly-launches-colab-pro.md/

6. Appendix

6.1 Code

```
def show_boxes(img2):
  """
  Find line boundaries on image text
  """
  from PIL import ImageDraw, Image
  with Image.open(img2) as img:
    drawing_object = ImageDraw.Draw(img)
    bounding_boxes = pageseg.segment(img.convert('1'), text_direction = 'horizontal-lr',
      black_colseps = False)['boxes']
    for box in bounding_boxes:
      drawing_object.rectangle(box, fill = None, outline = 'red')
    return img
```

6.2 Images

Figure A. Memory in Jupyter Hub Terminal

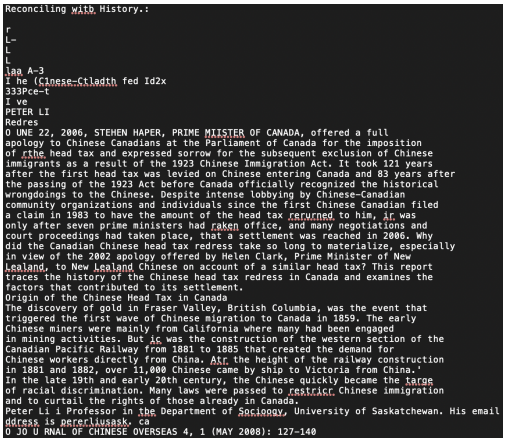


Figure B. txt File Output