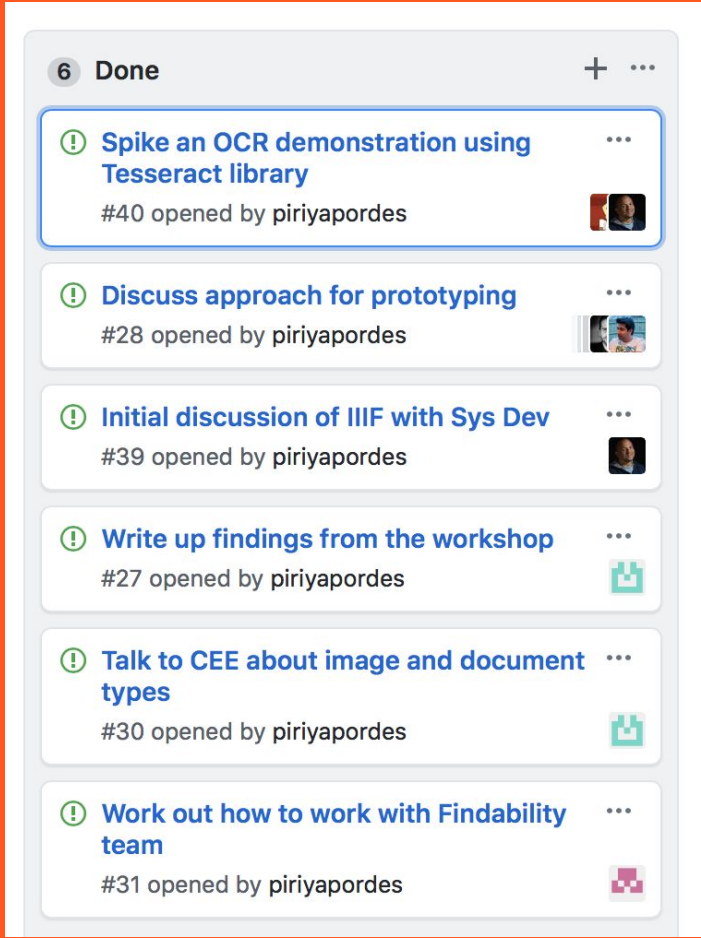
Experiments with text extraction

(Sprint 6)

The task


6 Done
Spike an OCR demonstration using Tesseract library
#40 opened by piriyapordes
Discuss approach for prototyping
#28 opened by piriyapordes
Initial discussion of IIIF with Sys Dev
#39 opened by piriyapordes
Write up findings from the workshop
#27 opened by piriyapordes
Talk to CEE about image and document types
#30 opened by piriyapordes
Work out how to work with Findability team
#31 opened by piriyapordes

So, what's a 'spike' exactly?

*"We carry out a Spike when we're about to embark on a new piece of work and are finding it hard to **make decisions about direction in business terms, technical terms, or both**. Similar in concept to a hackathon, a Spike is an investigatory piece of work which should:*

- *Answer **a single question***
- *Either be technical or customer-focused*
- *Reduce uncertainty and create a way forward*

*The key thing to note is that **a Spike doesn't directly contribute to an increment in the working software**. As a result, we don't estimate spikes, instead we timebox them."*

*Flewelling, P., 2018. **The Agile Developer's Handbook.** 1st ed. Pakt Publishing.*

Our ‘single question’

Can we quickly evidence the ability for OCR technologies to extract text from images for the purpose of **enhancing metadata, particularly for search?**

Not logged in Talk Contributions Create account Log in

WIKIPEDIA
The Free Encyclopedia

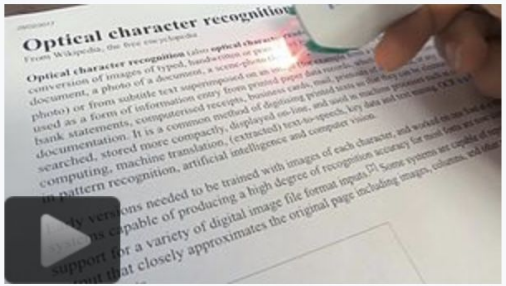
Article Talk Read Edit View history Search Wikipedia

Optical character recognition

From Wikipedia, the free encyclopedia

Optical character recognition (also **optical character reader**, **OCR**) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast).[1]

Widely used as a form of information entry from printed paper data records – whether



Video of the process of scanning and real-time optical character recognition (OCR) with a portable scanner.

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

OCR - with a focus on typed text

This was our first exposure to OCR technologies (*not our area of expertise)

Google Cloud

Vision API

Google Open Source

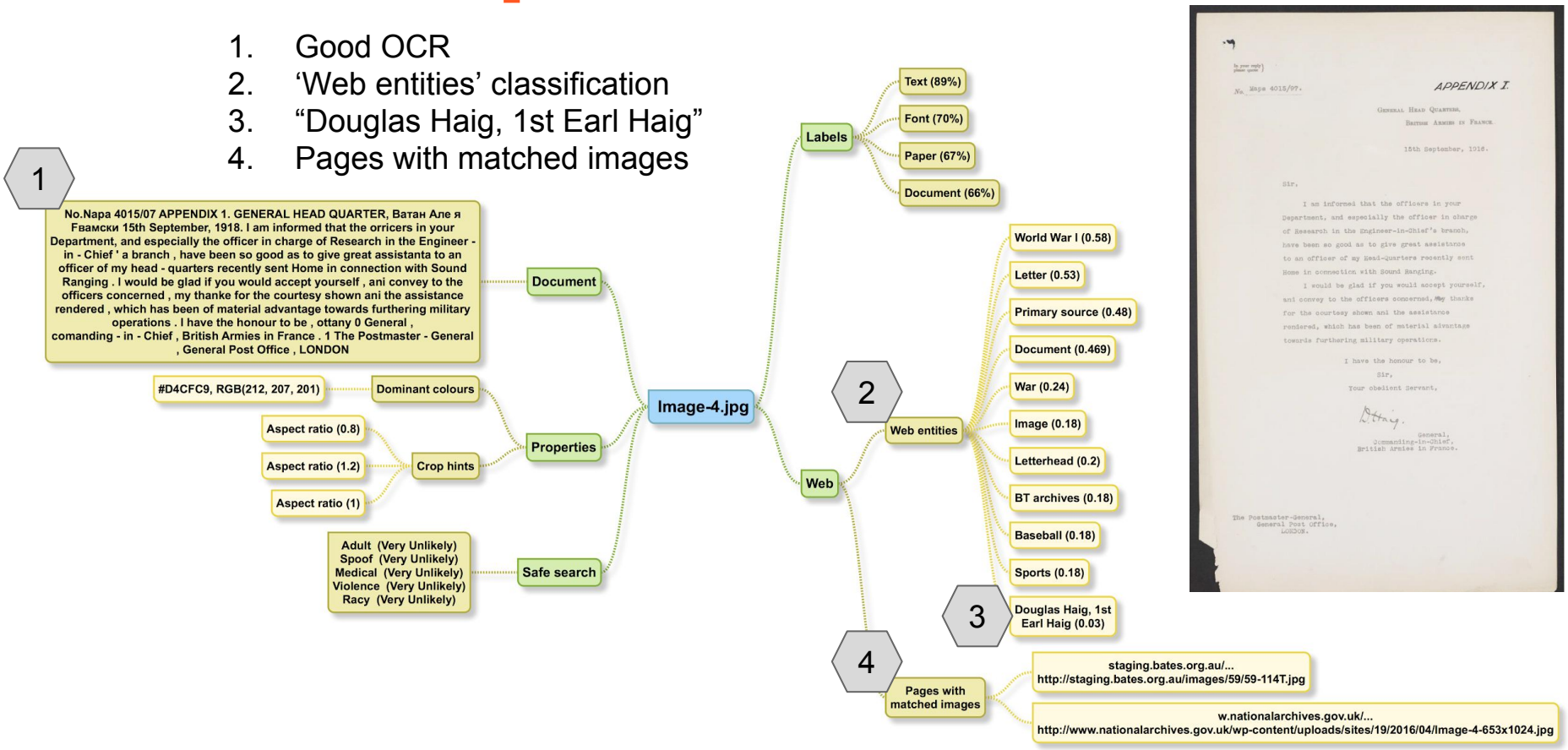
Tesseract OCR

*Tesseract engine was originally developed as proprietary software by HP Labs (starting in 1985). Released as Open Source in 2005. Sponsored by Google since 2006.

Google Cloud Vision API

Good OCR, and quite a bit more

1. Good OCR
2. ‘Web entities’ classification
3. “Douglas Haig, 1st Earl Haig”
4. Pages with matched images



APPENDIX I.

No. Napa 4015/07.

GENERAL HEAD QUARTERS,
BRITISH ARMIES IN FRANCE.

15th September, 1918.

Sir,

I am informed that the officers in your Department, and especially the officer in charge of Research in the Engineer-in-Chief's branch, have been so good as to give great assistance to an officer of my Head-quarters recently sent Home in connection with Sound Ranging.

I would be glad if you would accept yourself, and convey to the officers concerned, my thanks for the courtesy shown and the assistance rendered, which has been of material advantage towards furthering military operations.

I have the honour to be,
Sir,
Your obedient Servant,

Ottany.
General,
Commanding-in-Chief,
British Armies in France.

The Postmaster-General,
General Post Office,
LONDON.

An object

- image
  - Pages with matched images
    - http://www.nationalarchives.gov.uk/first-world-war/telecommunications-in-war/
    - https://www.telegraph.co.uk/technology/connecting-britain/battle-of-verdun/
    - http://www.westpol.co/wire/world-war-1-communication-wireless.html
    - http://www.westpol.co/wire/wireless-operator-world-war.html
    - http://www.westpol.co/wire/world-war-1-wireless-communication.html
    - http://www.pinsdaddy.com/world-war-ii-communication-devices_LqQu8Hn7jmqfGku8gLTeIQHLxXFR0E7%7C9fKoPBFTYEZ9vzRHokXVEJ%7CZ5wQpyyeKM3kJO18WwVatL3g6XU*BRA/
  - Fully Matched Images
    - http://www.nationalarchives.gov.uk/wp-content/uploads/sites/19/2016/04/TCB417E01723-739x1024.jpg
  - Partially Matched Images
    - https://www.telegraph.co.uk/content/dam/technology/Spark/btconnectingbritain/Gun-spotters-telephone.jpg?imwidth=480
  - 1
    - Document
      - RANGE SHORT OVER PRESS BUTTON WHILE SPEAKING SIGNALLING DEPRESS KEY HANDLES AS FOLLOWS OVER TO INDICATE RED ONCE MUCH OVER TWICE SHORT GREEN ONCE VERY SHORT TWICE RANGE WHITE ONCE 3 TIMES TO TELEPHONE
  - Properties
    - Dominant colors
      - #53535A, RGB(83, 83, 90)
    - Crop hints
      - 0.8
      - 1
      - 1.2
  - Safe search
    - Adult Very Unlikely
      Spoof Very Unlikely
      Medical Very Unlikely
      Violence Very Unlikely
      Racy Very Unlikely
  - 2
    - Labels
      - Technology
      - Electronics
      - Product
      - Hardware
      - Electronic component
      - Audio
      - Audio component
  - 3
    - Web entities
      - World War I
      - World War II
      - Early communication
      - World War
      - Telecommunications
      - Wireless
      - Telephone
      - BT Archives
      - War
      - BT Group
      - Mobile phones
      - Electrical telegraph
      - Communication

Tesseract OCR

Tesseract.js is a pure Javascript port of the popular Tesseract OCR engine.

This library supports **over 60 languages**, automatic text **orientation and script detection**, a simple interface for reading paragraph, word, and character **bounding boxes**. Tesseract.js can run either in a **browser** and on a server with **NodeJS**.

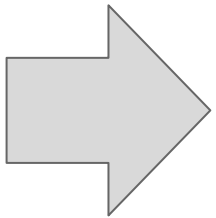
Check out the Example code and API docs on GitHub.

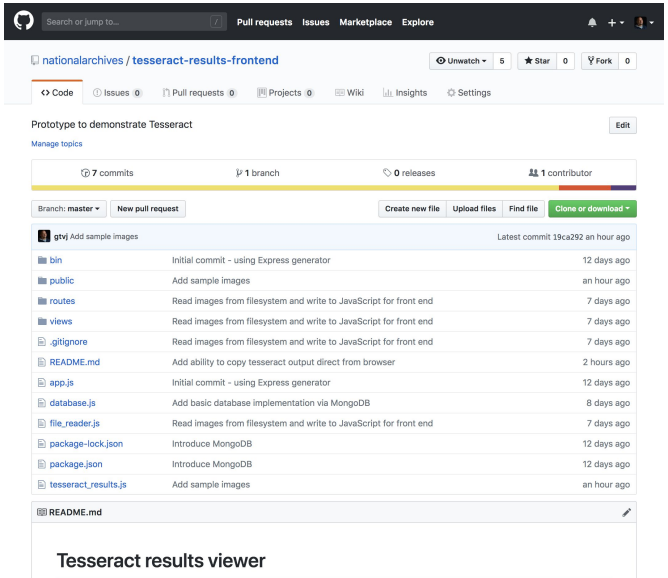
node.js

express

mongoDB

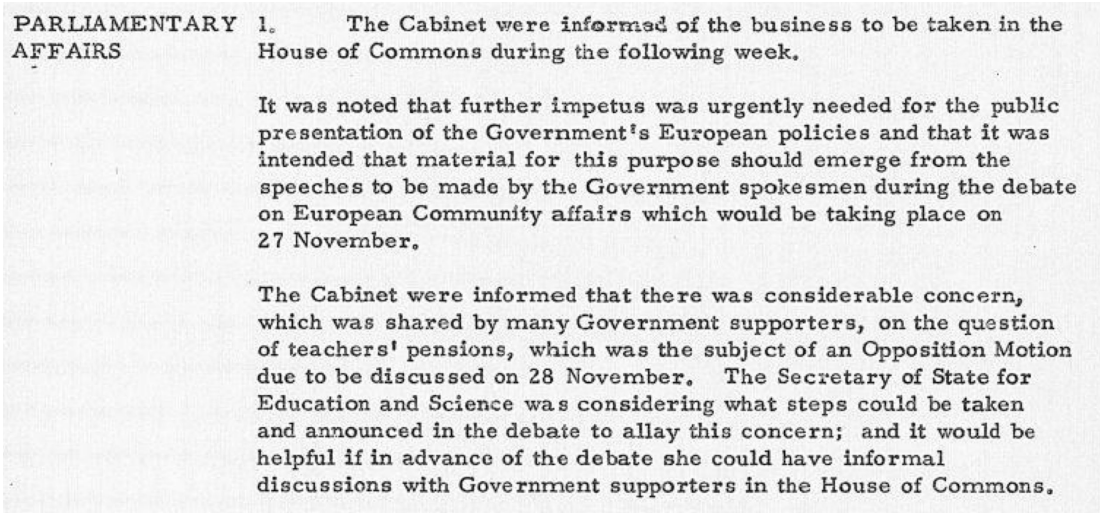
_.template()





https://github.com/nationalarchives/tesseract-results-frontend

Tesseract results ranged from very good...



PARLIAMENTARY AFFAIRS

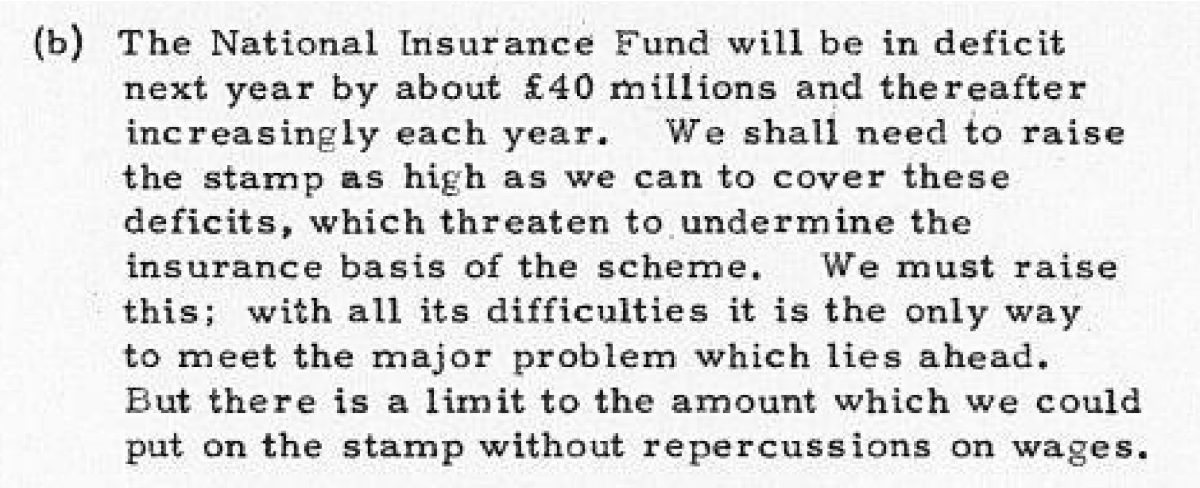
1. The Cabinet were informed of the business to be taken in the House of Commons during the following week.

It was noted that further impetus was urgently needed for the public presentation of the Government's European policies and that it was intended that material for this purpose should emerge from the speeches to be made by the Government spokesmen during the debate on European Community affairs which would be taking place on 27 November.

The Cabinet were informed that there was considerable concern, which was shared by many Government supporters, on the question of teachers' pensions, which was the subject of an Opposition Motion due to be discussed on 28 November. The Secretary of State for Education and Science was considering what steps could be taken and announced in the debate to allay this concern; and it would be helpful if in advance of the debate she could have informal discussions with Government supporters in the House of Commons.

PARLIAMENTARY 1c The Cabinet were informed of the business to be taken in the AFFAIRS Heuse of Commons during the following week. It was noted that further impetus was urgently needed for the public presentation of the Government's European policies and that it was intended that material for this purpose shoold emerge from the speeches to be made by the Government spokesmen during the debate on Eur0pean Community affairs which would be taking place on 27 November. . The Cabinet were informed that there was considerable concern, which was shared by many Government supporters, on the question of teachers' pensions, which was the subject of an OppositiOn Motion due to be discussed on 2.8 November. The Secretary of State for Education and Science was considering what steps could be taken and announced in the debate to allay this concern: and it w0uld be helpful if in advance of the debate she could have informal - discussions with Government supporters in the House of Commons.

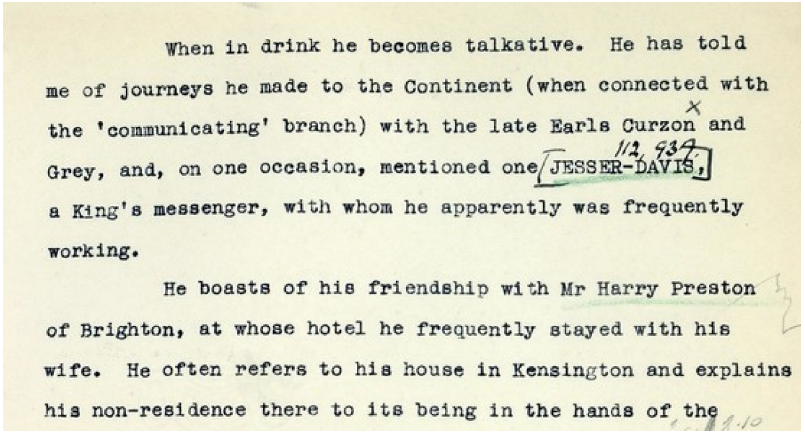
...to good



(b) The National Insurance Fund will be in deficit next year by about £40 millions and thereafter increasingly each year. We shall need to raise the stamp as high as we can to cover these deficits, which threaten to undermine the insurance basis of the scheme. We must raise this; with all its difficulties it is the only way to meet the major problem which lies ahead. But there is a limit to the amount which we could put on the stamp without repercussions on wages.

[h] The Natisnal Insurance Fund will be in deficit next year by sheet £41] millicns and thereafter increasingly each year. We shall nee-:1 te raise the stamp as high as we can tc ccver these deficits. which threaten tc undermine the insurance basis cf the scheme. We must raise this; with all. its difficulties it is the rent}: way te meet the majer prehlem which lies ahead. But there is a limit tn the arncnnt which we cculd Put an the stamp withcut repercussisns en wages

...to not so good



When in drink he becomes talkative. He has told me of journeys he made to the Continent (when connected with the 'communicating' branch) with the late Earls Curzon and Grey, and, on one occasion, mentioned one JESSER-DAVIS, a King's messenger, with whom he apparently was frequently working.

He boasts of his friendship with Mr Harry Preston of Brighton, at whose hotel he frequently stayed with his wife. He often refers to his house in Kensington and explains his non-residence there to its being in the hands of the

'H'hen in drink he heeenea talkative. I-le haa told me at" journeys he made to the continent [when connected with the 'cummunieating' branch} 1with the late Earle Cur-an: and Grey, and, on cue aceeeian, mentianed mefifiifiifirilfiafiggfl :1 King's messenger, with when he apparently wee fragment-1y working. He haaete of hie frienaehin with Mr Henry" Preaten of Brighton. at. 'lhDEIB hfltfil 11E frEQHEnt-ly starred with hie wife. He aft-in refer-a to hie lieuae in Kenaingtnn and explains his nan-residence there te ita being in the handle at the .-

Conclusions?

- Some **very promising results in good conditions:**
  - **Explore pre-optimization of images (https://scikit-image.org)**
  - **Explore using ML for confidence, training or even clean up**
- **The graph** might be interesting
  - Is there potential benefit in using OCR to expose relationships between records?
- We have only scratched the surface.