Historian's workflow

- Read documents in collection
- Collect interesting topics
- Snowball method:
  - Read again, collecting notes about selected topics
  - Add findings to “snowball”
  - Follow leads
  - Iterate
HistoRadar concept

Highlight places of potential interest in the historical document collection

- Extract information from text
- Radar shows points where information changes
  - Interesting places to start the “snowball”? 
- Example:
  - Opinion change: A-supports-B → A-opposes-B
HistoRadar concept

Highlight places of potential interest in the historical document collection

- Realistic first step
  - Track attendants to meetings of the British Cabinet
    - Who was suddenly absent?
    - Who re-appeared?
  - Named entities
    - Which countries start/stop being mentioned?
    - Which persons?
Source text acquisition
Source text acquisition


- PDF with OCR text
- Extraction of text
- Document splitting
WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS).

Minutes of a Meeting of the War Cabinet, held at 10, Downing Street, S.W., on Tuesday, 30th July, 1918, at 4.30 P.M.

Present:

The Right Hon. A. BONAR LAW, M.P. (in the Chair).

The Right Hon. W. M. HUGHES, Prime Minister of Australia.

The Right Hon. W. F. LLOYD, K.C., Prime Minister of Newfoundland.

Lieutenant-General the Right Hon. J. C. SMUTS, K.C.M.G., Minister for Defence, Union of South Africa.


The Right Hon. W. LONG, M.P., Secretary of State for the Colonies.

The Right Hon. E. S. MONTAGU, M.P., Secretary of State for War.

The Right Hon. LORD ROBERT CROFT, the Earl of Beaconsfield.

The following were also present:

Source text acquisition

SECRET

WAR CABINET.

WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), 480.
IMPERIAL WAR CABINET, 34.

Minutes of a Meeting of the War Cabinet and Imperial War Cabinet held at 10, Downing Street, S.W., on Tuesday, October 1, 1918, at 11.30 A.M.

Present:

The Right Hon. A. Bonar Law, M.P. (in the Chair).

The Right Hon. the Earl Curzon of Kedleston, K.G., G.C.S.I., G.C.I.E.
The Right Hon. G. N. Barnes, M.P.
The Right Hon. the Viscount Milner, G.C.M.G., Governor-General of South Africa.
The Right Hon. W. M. Hughes, Prime Minister of Australia.
The Right Hon. W. F. Lloyd, K.C., Prime Minister of Newfoundland.

Lieutenant-General the Right Hon. J. C. Smuts, K.C., Minister for Defence, Union of South Africa.
WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), IMPERIAL WAR CABINET, Minutes of a Meeting of the War Cabinet Street, S.W., on Tuesday, 34.

and Imperial War Cabinet held at 10, Downing October 1, 1918, at 1T30 A.M.

Present:
Source text acquisition

*iTfois Document is the Property of EIs Britannic Majesty^Goyernm^tX
Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.

WAR CABINET (WITH PRIME MINISTERS OF DOMINIONS), IMPERIAL WAR CABINET, ++Minutes of a Meeting of the War Cabinet Street, S.W., on Tuesday, 34.

and Imperial War Cabinet held at 1 0 , D o w n i n g October 1, 1918, at 1T30 A.M.
Present:

Minutes of a Meeting of the War Cabinet, Street, S.W., on Tuesday, 3rd and Imperial War Cabinet, held at 1:30 A.M. on October 1, 1918.

Present:

The Right Hon. A. BONAR LAW, M.P. (in the Chair).

The Right Hon. W. M. HUGHES, Prime Minister of Australia.

The Right Hon. G. N. BARNES, M.P., the Rt. Hon. W. F. LLOYD, KG., Prime Minister of Newfoundland.


SMUTS, KG., Minister for Defence, Union.

The Right Hon. the VISCOUNT MILNER, of South Africa, G.C.B., G.C.M.G., Secretary of State for War.

The Right Hon. W. LONG, M.P., Secretary of State for the Colonies.

The Right Hon. E. S. MONTAGU, M.P., Secretary of State for India.

Problem: several documents per file

Approach: find document start line, split there
Source text acquisition

*Itfois Document is the Property –of His Britannic Majesty’s Government*

Printed SECRET.
for the War Cabinet.
October
Li) 18.
WAR
CABINET.
S "J' 480.

patterns = [
    re.compile(r"\bthis\b.*\bdocument\b.*\bproperty\b", re.I),
    re.compile(r"\bdocument\b.*\bproperty\b.*\bhis\b +\britannic\b", re.I),
    re.compile(r"\bproperty\b.*\britannic\b +\bmajesy\b", re.I),
    re.compile(r"\bdocument\b.*\bproperty\b.*\bmajesy\b", re.I),
    re.compile(r"\bthis\b +\bdocument\b.*\bgovernment\b", re.I),
    re.compile(r"\bproperty\b +\bof\b.*\bgovernment\b", re.I),
]

Split if more than one pattern matches the line.
Header repeats in first pages of some documents

→ split only if document length > 60 lines
Document clean-up and date extraction
Document clean-up

• Biggest problem: words with spaces in them
• Regexp replacement: ( \b\S ) \b\s\b
• Unwanted side effect: single characters (like the article "a") concatenated to next word
• Use a word splitting library
Date extraction

• Which method?
  • Browse through provided NLP links: DANTE:
Meeting of the Cabinet to be held at No. 10, Downing Street, S.W.1., on WEDNESDAY, 23rd OCTOBER, 1936, at 11.0 a.m.

AGENDA. 1. FOREIGN AEF/-IR3. (9-) Preparations for the Five Power Conference. (Reference Cabinet 58 (36) Conclusion 3). Memoranda by the Secretary of State for Foreign Affairs. C.P. 268 (36) - already circulated. C.P. 278 (36) - already circulated. C.P. 284 (36) - to be circulated. ("b) The Situation- in Spain (if required).

(Reference Cabinet 58 (36) Conclusion 6).

The Cabinet had before them the following
Date extraction

• Which method?
  • Browse through provided NLP links: DANTE:
  • Problems: doesn't deal with "24 hours after 3 October" "between 5 and 7 October"
Date extraction

• What do we need all **dates** for?
  • most important is the date of cabinet meeting
    → extract "held on" date
  • we can do that with regular expressions
Date extraction

- Dates we have to deal with:

  Tuesday, 7th August, 1945, at 5-0 p.m. Street, S.W. 1,
  Thursday, 1st November, 1945, at 11 a.m.
  Tuesday, 1st January, 1946, at 11 a.m.
  December 9, 1916, at 11-30 A.M.
  March 1, 1917, at 11-30 A.M.
  Tuesday, June 5, 1917, at 11:30 A.M.
  Tuesday, January 1, 1918, at 11:30 A.M.
  Monday, April 1, 1918, at 11:30 A.M.
  Monday, July 1, 1918, at 12 noon.
  October 1, 1918, at 11:30 A.M.
  Thursday, 3, 1919, at 12
  Tuesday, July 1, 1919, at 11:30 A.M.
  Friday, June 8, 1917, at 11:30 a.m.
  Friday, August 15, 1919, at 11:30 A.M
  Friday, June 8, 1917, at 11:30 A.M.
  Tuesday, January 2, 1940, at 11 A.M
  WEDNESDAY, 21st JUNE, 1939, at 10:30 a.m
  MONDAY, 24th APRIL, 1939 at 5:30 p.m
  WEDNESDAY, 15th MARCH, 1939, at 11:0 a.m
  WEDNESDAY, 22nd MARCH, 1959, at 10:0 a.m
Date extraction

- Dates we have to deal with:

- Preprocessing (post-OCR):

  ```java
  pattern = Pattern.compile("(\b\S)\b\s\b");
  matcher = pattern.matcher(text);
  correctedText = matcher.replaceAll("$1");
  www.myregexp.com for Java
  ``

- Slot extraction:
  - year, month, day, time
  - day of week?
  - order of elements

- Remaining problems:
  - 1T30 A.M.
  - IE.30 a.m.
  - 10o30
Named Entity Recognition
Named Entity Recognition

• Need to extract named-entities to derive facts about them

• At the very least:
  • whether they are present
  • how many times in a document
Named Entity Recognition

- Three approaches:
  - Own regexp-based tagger
  - Stanford NER
  - OpenNLP NER

- Technical difficulties for compilation
  - Solved finally for OpenNLP
  - Likely similar for Stanford NER
Named Entity Recognition

- Adaptation to our Document.SegmentList:
  - OpenNLP tokenizer removes spaces
    - Span offsets do not match source text
  - Otherwise fine
  - Possible to use different libraries and compare
Cabinet meetings
attendant list extraction
Attendant list extraction

• List of attendants to meetings of the Cabinet
• Regular structure in documents
• List of attendants separated at beginning
  • labeled with words like "Present:"
  • finished with "1." or "]." (as OCR-error)
  • allows us to extract it with good recall even with relatively simple techniques
Attendant list extraction

- Approaches
  - Regular expressions
  - OpenNLP NER
- Structure of block elements not easy to parse
- Names variably denoted
  - Titles of honor
  - Position in the office
  - First name(s) and last name

Johannes, Souhail
Attendant list extraction

Example:

Major-General F. B. MAURICE, C.B.,
Admiral Sr. R. JELLICOE, GOB., O.M.,
Director of Military Office.
The Hon. SIR J. S. MESTON, K.C.S.L.,
G.O.V.O., First Sea Lord. The Hon. R. ROGERS, Minister of Public Works, Canada. The Hon. J. L. HAZEN, Minister of Marine, and Fisheries, and of the Naval Service, Canada.
Mr. H. O. M. LAMBERT, C. B., Colonial Lieutenant-Governor Provinces, India.
Implementation
Implementation

- Complete GUI application in Java
- Full source code available
- Free Software / Open Source: GPL v3
- Completed after this presentation
- Details in the final report
Implementation

- Attributed document segments
  - Begin and end character offsets
  - Arbitrary string attributes
- Analogous to other implementations
  - OpenNLP `opennlp.tools.util.Span`
  - GATE `gate.SimpleAnnotation`
- XML export of document with annotation
User Interface

Document

Radar (not yet implemented)
User Interface

HistoRadar View

File

Title: NO TITLE FOUND YET
Date: Friday, October 11, 1918, at 4 P.M
Date: 11 October 1918 at: p.m.
Date: For input string: ""
Identifier: cab-23-8-wc-484-5.txt

Source Document

The Western Front.

The Chief of the Imperial General Staff reported that a telephone message had been received from General Du Cane, who said that the Germans appeared to be falling back along the whole line between the Ailette and the Meuse. Our troops were following. 2. The First Sea Lord reported that the Irish mail boat "Leinster," outward bound from Kingstown, had been torpedoed and sunk on the previous morning. Two torpedoes had been fired, the first at 9:28, and the second at 9:47, and the vessel had sunk three minutes later. A very heavy sea was running at the time, which had increased the difficulty of rescue work. It was feared that 50 officers and 300 other ranks had been lost.

Steamship "Leinster."


3. The Prime Minister, dealing with the Conferences which had been held in Paris between the 5th October and 9th October, 1918, said that he went over to confer with the French and Italian Governments, after the discussions which had taken place in the War Cabinet, mainly with regard to the situation in Turkey and Bulgaria, but, after arrival in France, the news was published of the German proposal of an armistice in order to negotiate peace, on the basis of President Wilson's 'Fourteen Points.' For some days no official intimation was received by any of the Prime Ministers, either from the United States of America or from any neutral country, as to this proposed armistice. and, although it appeared in the newspapers that Prince Max of Baden had read out to the Reichstag the message which he had despatched to President Wilson, it was not possible for some days, owing to lack of official information, for the Prime Ministers in Paris to deal with the question. The
User Interface

HistoRadar View

Title: NO TITLE FOUND YET
Date: Friday, October 11, 1918, at 4 P.M
Date: 11 October 1918 at: 4 p.m.
Date: For input string: ""
Identifier: cab-238-wc-484-5.txt

XML source of the annotated document

The text of the document is as follows:

The Chief of the Imperial General Staff reported that an telephone message had been received from General Du Cane, who said that the Germans appeared to be falling back along the whole line between the Ailette and the Meuse. Our troops were following. 2. The First Sea Lord reported that the Irish mailboat "Leinster," outward bound from Kingstown, had been torpedoed and sunk on the previous morning. Two torpedoes had been fired, the first at 928 and the second at 947, and the vessel had sunk three minutes later. A heavy gun was running at the time, which had increased the difficulty of rescue work. It was feared that 50 officers and 300 other ranks had been lost.

Steamship "Leinster."


3. The Prime Minister, dealing with the Conferences which had been held in Paris between the 5th October and 9th October, 1918, said that he went over to confer with the French and Italian Governments, after the discussions which had taken place in the War Cabinet, mainly with regard to the situation in Turkey and Bulgaria; but, after arrival in France, the news was published of the German proposal of an armistice in order to negotiate peace, on the basis of President Wilson's "Fourteen Points." For some days no official intimation was received by any of the Prime Ministers, either from the United States of America or from any neutral country, as to this proposed armistice, and, although it appeared in the newspapers that Prince Max of Baden had read out to the Reichstag the message which he had despatched to President Wilson, it was not possible for some days, owing to lack of official information, for the Prime Ministers in Paris to deal with the question. The representatives of the three Governments, however, met every day and discussed the situation, and the Military Representatives at Versailles, and, as a preliminary step, directed their attention to the terms of an armistice. All the military authorities demanded very drastic terms if an armistice were to be granted, and the Military
User Interface

• Ideas for additional features
  • Support “snowball” method
    – One-click bookmarking
      • Select text in document if desired
      • Click bookmark button
      • Bookmark added with optional citation text and notes
      • Export to HTML, BibTeX, Zotero, …
    – Display area for “snowball” bookmarks
  • Integrated search with query expansion
Questions?

http://historadar.googlecode.com/
Image sources

- Snowballing

  - http://www.flickr.com/photos/artsmonkey/3250602627/